

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМ. ІГОРЯ СІКОРСЬКОГО ”
Факультет електроніки
Кафедра електронної інженерії

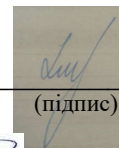
До захисту допущено
Завідувач кафедри
В. І. Тимофєєв
“ ” 20__р.

Дипломна робота

освітнього рівня «бакалавр»
за спеціальністю 153 мікро- та наносистемна техніка

на тему Аналіз усереднених електрокардіограм у нормі методами машинного
навчання

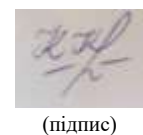
Виконав студент 4 курсу, групи ДМ-62
Фогель Дмитро Андрійович
(прізвище, ім'я, по батькові)


(підпис)

Керівник доц. каф. ЕІ, доц., к.т.н. Попов А.О.
(посада, вчене звання, науковий ступінь, прізвище та ініціали)


(підпис)

Рецензент доц. каф. ЕПС, доц, к.т.н. Клен К.С.
(посада, вчене звання, науковий ступінь, прізвище та ініціали)


(підпис)

Засвідчую, що у цій дипломній роботі немає
запозичень з праць інших авторів без
відповідних посилань.

Студент 
(підпис)

Київ - 2020 року

**Національний технічний університет України
“Київський політехнічний інститут імені Ігоря Сікорського”**

Факультет електроніки
Кафедра електронної інженерії
Освітній рівень «бакалавр»
за спеціальністю 153 мікро- та наносистемна техніка

ЗАТВЕРДЖУЮ

Завідувач кафедри

В. І. Тимофєєв

“ ____ ” ____ 20__ р.

**З А В Д А Н Н Я
НА ДИПЛОМНУ РОБОТУ СТУДЕНТУ**

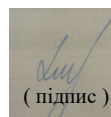
Фогелю Дмитру Андрійовичу

1. Тема роботи: Аналіз усереднених електрокардіограм у нормі методами машинного навчання
керівник роботи Попов Антон Олександрович, к.т.н., доц.
затверджені наказом по університету від “ ____ ” ____ 20__ року № ____
2. Строк подання студентом роботи 12.06.2020
3. Вихідні дані до роботи: база даних для аналізу ЕКГ здорових суб'єктів.
4. Зміст дипломної роботи (перелік питань, які потрібно розробити): розглянути методи і засоби стандартної електрографії, методи отримання усередненої електрокардіограми, методи машинного навчання, обрати метод машинного навчання для визначення аномалій, опрацювати базу даних, написати код робочої програми для пошуку аномалій раніше розглянутим та обраним методом, зробити висновки по роботі.
5. Перелік графічного (ілюстративного) матеріалу (із зазначенням обов'язкових креслень, плакатів, презентацій тощо): рисунки в тексті пояснювальної записки, ілюстрації до захисту.
6. Дата видачі завдання: 01.12.2019

КАЛЕНДАРНИЙ ПЛАН

№ з/П	Назва етапів виконання дипломної роботи	Строк виконання етапів роботи	Примітка
1	Огляд літератури з методів стандартної електрокардіографії	Грудень 2019 – лютий 2020 року	
2	Огляд літератури про SAECSG	Лютий 2020 року	
3	Огляд літератури про методи машинного навчання	Лютий – березень 2020 року	
4	Огляд програми Cardiol	Березень 2020 року	
5	Обробка вхідних даних	Березень – квітень 2020 року	
6	Написання оглядових розділів	Квітень 2020 року	
7	Створення коду для обробки даних та проведення експерименту	Травень 2020 року	
8	Обґрунтування отриманих результатів	Травень 2020 року	
9	Підготовка дипломної роботи до захисту	Червень 2020 року	

Студент


 (підпис)

Фогель Д.А.
 (прізвище та ініціали)

Керівник роботи



Попов А.О.

ЗМІСТ

ВСТУП	8
1 МЕТОДИ І ЗАСОБИ ЕЛЕКТРОКАРДІОГРАФІЇ	10
1.1 Стандартна електрокардіографія	10
1.2 Задачі аналізу ЕКГ	15
1.3 Реєстрація ЕКГ	17
1.4 Приклад стандартного кардіографу	23
1.5 Приклади нестандартних методів отримання ЕКГ	25
1.6 Усереднений сигнал ЕКГ	28
1.7 Постановка задачі роботи	35
1.8 Висновки до розділу 1	36
2 МЕТОДИ МАШИННОГО НАВЧАННЯ В АНАЛІЗІ SAECG	37
2.1 Штучний інтелект, машинне навчання, методи та алгоритми машинного навчання	37
2.2 Застосування методів машинного навчання в аналізі ЕКГ	42
2.3 Вибір методу для використання в даній роботі	45
2.4 Висновки до розділу 2	53
3 ЕКСПЕРИМЕНТАЛЬНА ЧАСТИНА	54
3.1 Опис даних	54
3.2 Постановка експерименту	57
3.3 Результати	68
3.4 Висновки до розділу 3	79
ВИСНОВКИ	80
ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ	82
Додаток А	85
Додаток Б	87

РЕФЕРАТ

Дипломна робота: 88 с., 3 ч., 10 табл., 56 рис., 2 дод., 27 джерел.

ЕКГ, УСЕРЕДНЕНИЙ КАРДІОКОМПЛЕКС, МАШИННЕ НАВЧАННЯ, АНАЛІЗ ПАРАМЕТРІВ ЕКГ, ПОШУК АНОМАЛІЙ

Метою дипломної роботи є визначення різниці між пацієнтами з нормальними сигналами електрокардіограми. Для цього необхідно створити норму для групи молодих людей та визначити їх «відстань» до створеної норми. На основі отриманих результатів складені висновки про різницю між клінічно здоровими та схожими пацієнтами.

В першому розділі розглянуті теоретичні відомості про історію ЕКГ, методику отримання сигналу ЕКГ, способи отримання усередненого комплексу та можливе їх використання.

В другому розділі описується теорія машинного навчання, запропонований метод вирішення поставленого перед роботою завдання. Відповідно запропонованого способу розглянуті доступні алгоритми машинного навчання та обраний найбільш підходящий (Isolation Forest).

В третьому розділі описані використовувані вхідні дані. Вони отримані з бази даних курсантів військового училища. Вікова категорія – 17 - 19 років, всі пацієнти вважаються клінічно здоровими. На стадії формування завдання дипломної роботи було прийнято рішення використовувати усереднений кардіокомплекс. В практичній частині створений код для вичитування і форматування вхідних даних з файлів. Використовуючи обраний метод побудована область норми та визначені аномалії відносно неї. Отримані результати порівняні з аналізом програми Cardiol. Всі пацієнти, судячи з результатів програми, мають деякі відхилення в комплексі показників, що відповідають за стан міокарду в різній мірі. Зроблено висновок про поняття норми для різни вікових категорій пацієнтів та запропоновані методи покращення результатів експерименту.

ABSTRACT

Bachelor's thesis: 88p, 3 p., 10 tables, 46 figures, 2 app, 27 references.

ECG, SIGNAL-AVERAGED ELECTROCARDIOGRAM, MASHINE LEARNING, ECG PARAMETERS ANALYSIS, OUTLIER DETECTION, ISOLATION FOREST

The purpose of this thesis is determination the difference between patients with normal ECG signals. Necessary to create a “normal cloud” for a group of young people and determine their “distance” to the established “cloud”. Based on the results obtained, it is possible to draw conclusions about difference between clinically healthy and similar patients. Partially answer the question: “How healthy is the patient?”.

In the first section, there are theoretical information about the history of the ECG, method of obtaining an ECG signal, methods of obtaining an averaged ECG complex and their possible using.

In the second section there are describes the theory of machine learning, proposed method of increasing main problem of labor. According to the proposed method, the available machine learning algorithms are considered and the most suitable (Isolation Forest) was selected. It is advantages are:

1. The possibility to work with multidimensional input data;
2. Low consumption computer resource;
3. The possibility of automatic formation of a “normal” sample.

The third section describes the inputs used. They are obtained from the database of cadets of military school. Age category – 17-19 years, all of patients are clinically healthy. At the stage of forming the thesis task, it was decided to use an averaged cardio complex. In the practical part, code is created for reading and formatting input data from files. The “norm cloud” was constructed and anomalies are defined. The obtained results was

comparable with the analysis of the program, that names Cardiol. All patients have some deviations in the set of parameters that are responsible for the state of the myocardium. Interesting, that 36% of them do not belong to a given age category. They are 67% of similar records in the entire database. Based on this fact conclusion – concept of norm for different age categories of patients is differ. The following steps have been proposed to improve the outcomes:

1. Refinement and increase the database;
2. To find the most weighty parameters for used algorithm;
3. More correctly find the parameters.

ВСТУП

«Здоров'я – це дорогоцінність, і притому єдина, заради якої варто не тільки не шкодувати часу, сил, праці і всяких благ, але і жертвувати заради нього часткою самого життя, оскільки життя без нього стає нестерпним і принизливим» (М. Монтень).

Щоб зберегти своє здоров'я необхідно знати себе, досліджувати свій організм і вносити поправки не лише тоді, коли щось зламається. Тіло потребує деякого «обслуговування» завжди, а щоб правильно обслуговувати, необхідно знати що відбувається в середині. Саме тому медицина розвивається не лише в напрямку вирішення проблем, а й їх попередження. Для цього ми маємо досліджувати величину здоров'я кожного окремого індивіду, щоб правильно використовувати превентивні дії, які допоможуть уникнути або відтягнути момент настання хвороби, щоб мати запас часу для підготовки.

Будь-яка частина організму потребує догляду, в тому числі і серце. Доволі велика кількість людей має проблеми з серцево-судинною системою, і ще більша навіть не уявляють, що їх може чекати завтра, якщо не приділити цьому питанню час сьогодні.

Дійсно, важливо знати про стан свого серця, але лікар не завжди може приділити достатньо уваги вашій цікавості, адже у нього можуть бути на розгляді випадки сильних, можливо, навіть смертельних, відхилень. Тому необхідно забезпечити деякий автоматичний аналіз показників систем організму. Тобто створити програмне забезпечення, яке здатне оцінити стан Вашого серця за короткий проміжок часу, що дозволить фахівцю дати деякі рекомендації для підкріплення здоров'я.

Найрозповсюдженішим та найдоступнішим методом аналізу серцево-судинної системи є використання електрокардіограми. В деяких випадках

використовується усереднена електрокардіограма та її параметри. Метою та завданнями даної роботи будуть :

1. Розглянути методи отримання сигналу електрокардіограми, його усереднений вид та параметри;
2. Зрозуміти, яким чином та для чого використовують отримані дані;
3. Запропонувати метод аналізу отриманих сигналів та їх параметрів;
4. Розглянути теорію нейронних мереж, методів машинного навчання;
5. Обрати метод машинного навчання для аналізу усереднених електрокардіограм та їх параметрів;
6. Створити програму, що визначає «величину здоров'я серця» на основі наявних даних, можливо знайти відхилення від норми попередньо клінічно здорового пацієнта;
7. Зробити висновки про отримані результати, запропонувати покращення отриманого методу або інший метод для розв'язання поставленої задачі.

1 МЕТОДИ І ЗАСОБИ ЕЛЕКТРОКАРДІОГРАФІЇ

1.1 Стандартна електрокардіографія

Електрокардіографія – це методика реєстрації та дослідження електричних полів, які виникають під час роботи серця. Методика базується на отриманні запису різниці потенціалів на поверхні тіла, виникаючих в результаті роботи серця – електрокардіограми (ЕКГ). Важливо використовувати електрокардіографію не як окремий метод, а в комплексі, на рівні з іншими методами, хоча він являється основним в питаннях дослідження порушень ритму серця.

Серце відповідає за рух крові по організму людини. За рахунок цього руху всі клітини організму мають джерело живлення киснем та різними речовинами. Серце – чотирикамерне, має повноцінну перегородку між шлуночками і передсерддями. Це означає, що кров не змішується в серці і може розділитися на повноцінні «ізольовані» контури.

Існує два кола кровообігу: мале і велике. Судини малого кола проводять бідну киснем кров з серця до легень, а збагачену назад до серця. По великому колу збагачена кров рухається по організму від серця, а повертається збіднена. Повний цикл виглядає так: з правого передсердя, під час скорочення, збіднена кров потрапляє у правий шлуночок, звідти по малому колу кровообігу до легень, збагачена киснем в легенях вона повертається до серця, а саме до лівого передсердя, з нього до правого шлуночка, після скорочення якого насичена кров розходить по організму, обмінюється речовинами з органами і повертається до правого передсердя (рисунок 1.1). Варто зауважити, що скорочення передсердь, як і шлуночків, виконується попарно одночасно (спочатку передсердя, потім шлуночки). Повний цикл протікання крові по організму триває 19–24 секунди.

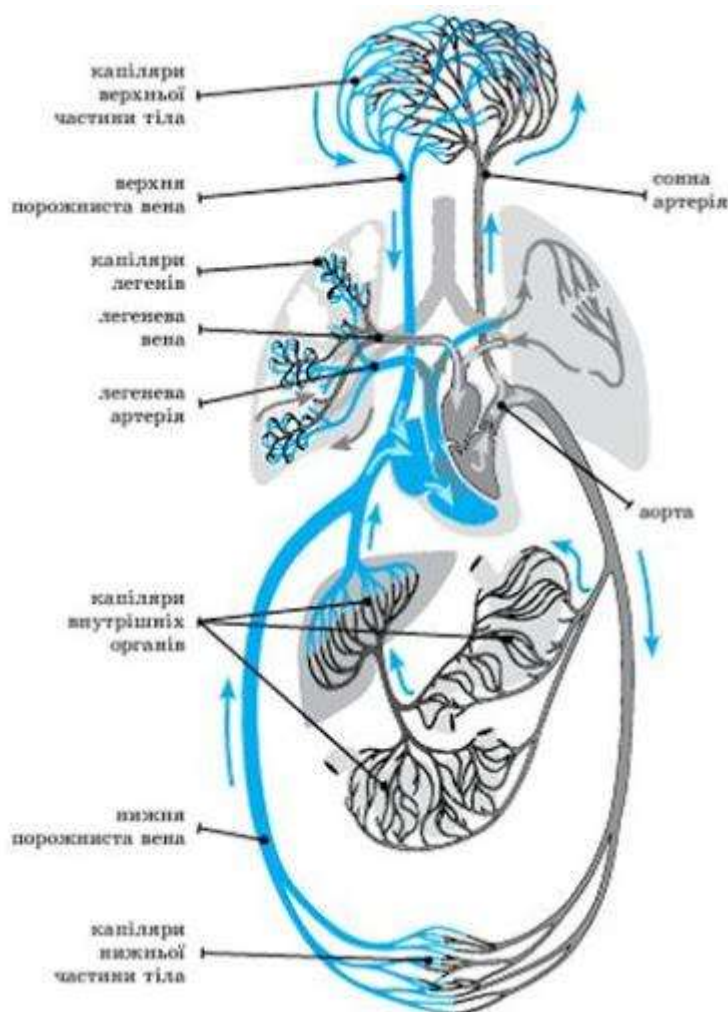


Рисунок 1.1 – Схематичне зображення кіл кровообігу людини

В кожному серцевому циклі електричне збудження міокарду починається в спеціальній зоні правого передсердя, яке називається синоатриальний вузол. Далі деполяризація розходить по волокнам міокарда передсердь, досягає іншої зони в нижній частині передсердь – атріовентрикулярний вузол, де затримується на деякий час, порядку 200 мкс. Далі електричне збудження рухається по пучку Гіса, а потім по більш повільним структурам – волокнам Пуркінє (рисунок 1.2).

Провідна система серця

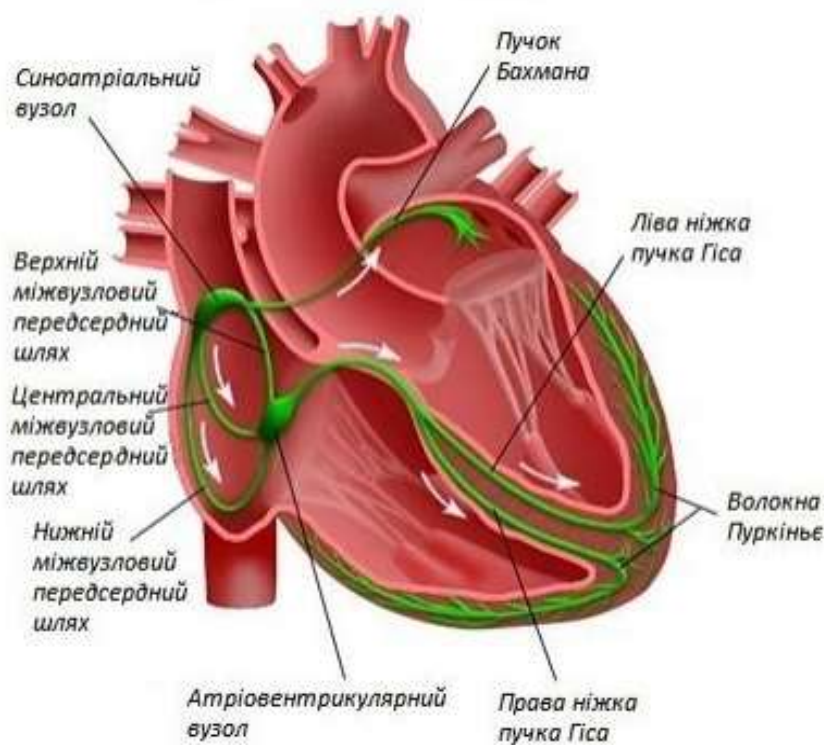


Рисунок 1.2 – Провідна система серця

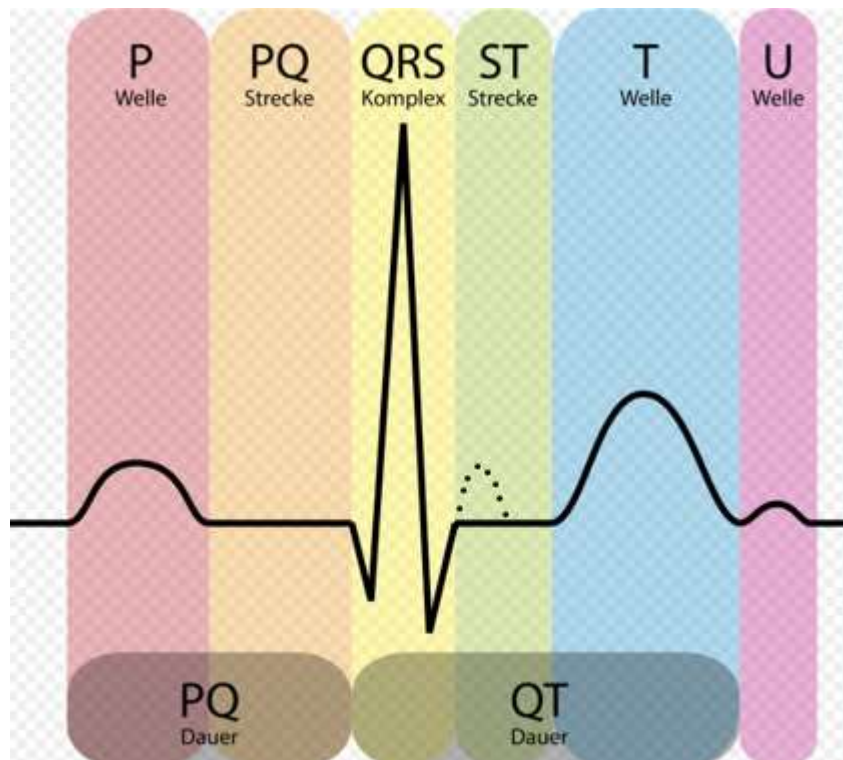


Рисунок 1.3 – Нормальний кардіокомплекс

Скорочення будь – якого м'язу протікає з електричними змінами, а значить, що реєстрація ЕКГ потребує певного відмежування сигналів скорочення серцевого м'язу від інших. Найпростіший спосіб – помістити пацієнта у стан максимального фізичного спокою, щоб інші м'язи не скорочувались, викликаючи шум та сторонні сигнали. Важливо те, що реєструється не саме скорочення, а збудження, що його викликає. Кожне збудження характеризує диполь, який характеризується дипольним вектором різної величини і напрямку. Інтегральний вектор – результат алгебраїчної сумації різних векторів, які виникають від кожної клітини м'язу. Цикл серцевої діяльності можна розбити на п'ять основних зубців, які мають назви Р, Q, R, S, Т (рисунок 1.3). Позначення не несуть якогось смислового навантаження, ці букви були обрані піонерами електрокардіографії випадково, коли метод почав розвиватися.

Об'єм міокарда передсердь набагато менший за міокард шлуночків, тому електричні зміни, пов'язані зі скороченням передсердь, менші.

Зубець Р виникає на кардіограмі в результаті збудження передсердь. При збудженні всіх відділів різниця потенціалів тимчасово зникає, оскільки потенціал дії знаходиться на стадії «плато». Збудження шлуночків починається при реполяризації стінки і утворює зубець Q. Вектор змінює напрям на протилежний. Далі відображається розповсюдження збудження від ендокарду до перикарду (зубець R). За ним настає остання реполяризація (S). QRS відображає розповсюдження в шлуночках і співпадає з реполяризацією передсердь. Коли шлуночки повністю охоплені збудженням різниця потенціалів деяких точок зникає. Зубець Т – найбільш змінна частина ЕКГ. Наприклад при інфаркті міокарда він приймає інверсний стан. На рисунку 1.3 зображено шість зубців, хоча раніше було зазначено, що їх п'ять. Останній зубець U є не завжди. Він вказує на остаточну реполяризацію волокон Пуркінє. Амплітуда зубця R може досягати амплітуди порядку 0.5 – 1 мВ. Також бачимо зубець, позначений крапками, це зубець J, або, так звана, хвиля Осборна, що виникає з проблемами, пов'язаними з реполяризацією шлуночків.

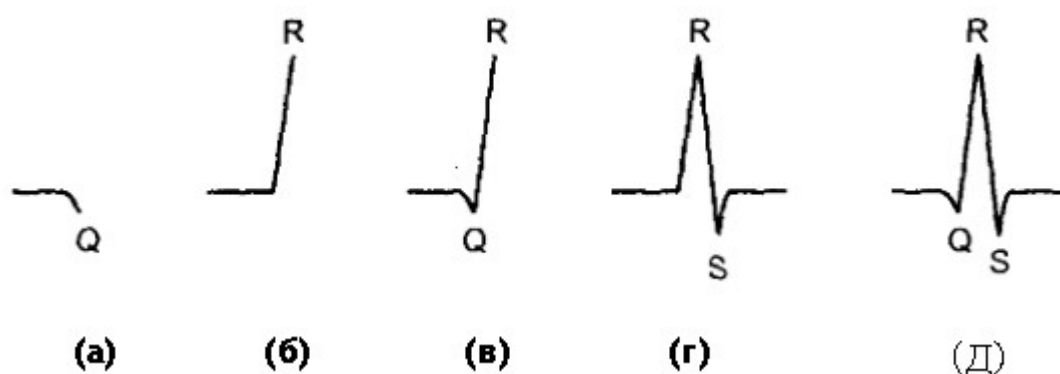


Рисунок 1.4 – Складові комплексу QRS

Зубці Q, R, S разом складають QRS комплекс, інтервал між закінченням зубця S та початком T відомий, як сегмент ST. Частини комплексу QRS називаються так, як представлено на рисунку 1.4. Якщо перший зубець від'ємний, то його позначають Q (рисунок 1.4.а). Всі зубці, напрямлені вгору позначають R (рисунок 1.4.в), незалежно від того, є переднім зубець Q чи ні (рисунок 1.4.б). Будь-який від'ємний зубець комплексу QRS після зубця R позначається буквою S (рисунок 1.4.д), незалежно від того, був раніше Q чи ні (рисунок 1.4.г). В середньому висота R коливається в межах 0.3 – 1 мВ.

На практиці, аналіз ЕКГ – пошук деяких шаблонів, порівняння отриманого сигналу з відповідними для деяких патологій. Тобто лікар, аналізуючи кардіограму пацієнта керується принципом прецеденту, шукає те, що бачив раніше у інших хворих, або в підручнику з кардіології [1]. До списку патологій відносяться:

1. Атріовентрикулярна блокада 1-го, 2-го, 3-го ступеня;
2. Блокада правої або лівої ніжки пучка Гіса (видно на рисунку 1.2);
3. Різкі відхилення електричної осі серця;
4. Фібриляції передсердь або шлуночків;
5. Тахікардія;
6. Аритмія;
7. Синдром Вольфа Паркінсона – Уайта (WPW);

8. Та інші.

Розпізнавання цих патологій і є основною задачею аналізу ЕКГ на сьогоднішній день.

1.2 Задачі аналізу ЕКГ

Щорічно в Україні вмирає більше 400000 людей від серцево-судинних захворювань та їх місце займають інші люди, цим вони не лише підтримують гірку статистику, а й сприяють її росту по причині несвоєчасного визначення ризиків та проблем. Хворі не з'являються просто так. Місце померлих хворих займають ті, кого раніше вважали здоровим. Тому необхідно визначитись з тим кому і навіщо необхідна ЕКГ.

Вона необхідна хворому, оскільки вказує на патологію, можливі причини та методи її подолання. Теоретично, ЕКГ необхідна і здоровій людині. Але на практиці це скоріше виняток, ніж правило. Фраза про відсутність ЕКГ патологій гріє думку і сприймається як «Ви здорові». Але наскільки здорові? Скільки це ще буде продовжуватись? Чи необхідно задумуватись про зміцнення серцево – судинної системи? Яким чином це робити та якої профілактики необхідно дотримуватись?

Сучасний алгоритм оцінки ЕКГ складається з пошуку патологій, а не оцінки стану серцевої діяльності. Такий алгоритм не відповідає інтересу клінічно здорової людини, адже не може говорити про рівень здоров'я серця та судинної системи, ризики виникнення патологій та які його функціональні межі. Лікарі не аналізують стан серцево-судинної системи, якщо пацієнт клінічно здоровий, вони на все можуть відповісти – на ЕКГ патологій не виявлено. Тому все частіше починають говорити про те, що необхідно використовувати програмне забезпечення в аналізі ЕКГ. Воно дозволить швидко і досить точно визначити функціональні показники,

ризика захворювання та вказати на методи покращення стану серцево – судинної системи.

Дослід [2] показує, що клінічно здорові пацієнти мають деякі відхилення від норми, які не були помічені при перегляді сигналів лікарем. До того ж використаний алгоритм не несе трудових витрат лікаря для дослідження та розрахунків. Приведений алгоритм враховує наступні показники: частота серцевих скорочень (ЧСС), R-R, P-Q, P-Qс, Q-Тс, Р, QRS, індекс PQs, індекс фази плато, індекс фази відпочинку міокарда, індекс фізіологічності систоли серця. На основі цих даних складається загальна картина функціонування серця, ризику виникнення тахікардії та рекомендовані показники ЧСС при фізичних навантаженнях. Також можливий аналіз варіабельності ритму серця для виявлення деяких показників, але потребує частіших вимірів та створення таблиці серцевого ритму.

Введення програмного забезпечення для первинного аналізу може значно зменшити час аналізу та швидко вказати на слабкі місця серцево – судинної системи пацієнта. Головне завдання – створення програмного забезпечення яке виконує первинний аналіз з максимальною точністю.

Основний мотив статті [3] – електрокардіографія, зрозуміла користувачеві. ЕКГ на даний момент найдешевший і найрозповсюдженіший метод дослідження активності серцевого м'язу. Кардіографи стають меншими, а аналіз вийшов за межі кабінету лікаря. В тексті представлена програма, яка дозволяє швидко оцінювати показники роботи серця, використовуючи, навіть короткі записи. Враховуючи те, що дане програмне забезпечення може знаходитися на віддаленому сервері, так званій «хмарі», моніторинг власного серця стає доступним в будь-якій точці світу, за наявності покриття мережі Інтернет. Не меншою перевагою є більша «прискіпливість» комп'ютера до непомітних чи неоднозначних для людини випадків. Автор говорить про «мінімальні зміни, які не помітні звичайному спостерігачеві»

Інтерфейс виконаний в максимально зручному і інтуїтивно зрозумілому стилі «traffic light logic» (рисунок 1.5).

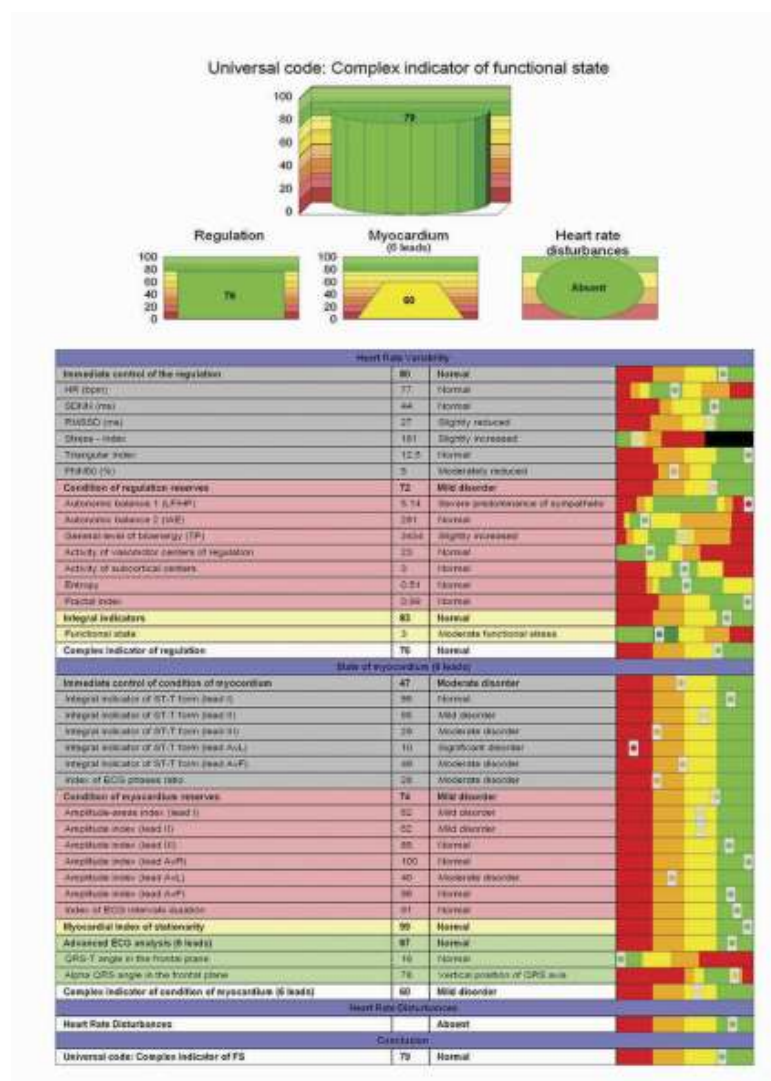


Рисунок 1.5 – «traffic light logic» в системі підрахунку балів стану серцевого м'язу

1.3 Реєстрація ЕКГ

Перший кардіосигнал, прототип сучасної ЕКГ був зареєстрований В. Ейнтховеном в 1912 році. Це був феноменальний прорив у сфері вивчення організму, а саме серця. В 1924 році автору методу було вручено Нобелівську премію по медицині. Але цій події сприяла довга робота, виконувана не одним Віллемом Ейнтховеном.

Ще в 1842 році Дюбуа-Реймонд помітив виникнення електричного струму під час скорочення серця жаби. Цей струм було названо потенціалом дії. Спостерігати його зміни стало можливим лише через 30 років. Ліппман винайшов ртутний капілярний електрометр. І з допомогою мікроскопа стало можливо побачити та записати зміни струму потенціалу дії. Огуст Валлер вперше записав ці потенціали у людини.

В 1887 році Ейнтховен демонструє на першому міжнародному конгресі фізіологів в Лондоні доволі чітку криву потенціалів, які виникли в наслідок діяльності серця. Вона була виміряна з допомогою струнного гальванометра. Через 2 роки Віллем Ейнтховен назвав даний сигнал електрокардіограмою. В 1901 році він конструює апарат на основі свого гальванометра, який був здатен записувати ЕКГ. Вага даного електрокардіографа була понад 100 кг.

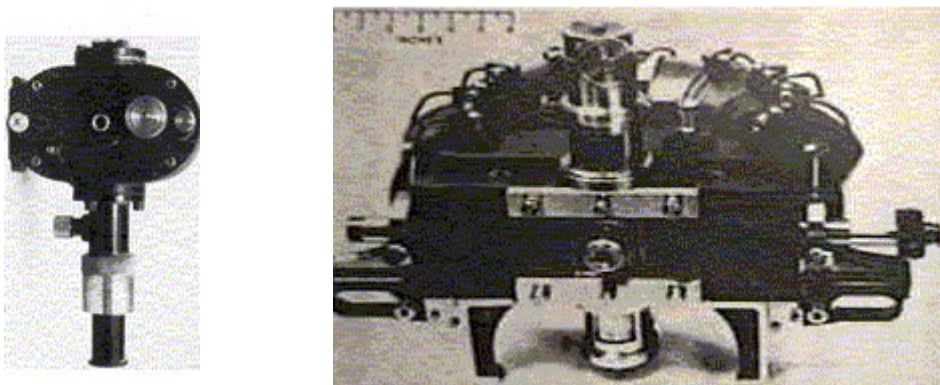


Рисунок 1.6 – Струнний гальванометр та перший апарат ЕКГ на його основі
1901 р

В 1906 році ним була запропонована методика опису векторного аналізу кардіограми в рівнобедреному трикутнику. Через 6 років остаточно був сформований принцип зняття 3-канального сигналу.



Рисунок 1.7 – Ейнтховен записує свою ЕКГ.

Як видно з рисунку 1.7 Віллем Ейнтховен занурював руки у спеціальні резервуари. Це так звані «мокрі» або «вологі» електроди. Між провідником і поверхнею тіла має бути максимально надійний контакт, тому Ейнтховен брав резервуари, набирав в них фізіологічний розчин, який добре проводить струм та занурював в нього електроди. Таким чином виникає максимально надійний контакт з поверхнею тіла. До того ж зі шкіри відразу змивались усі жирові плівки, які могли бути перешкодою для сигналу. Навіть зараз, через більш як сотню років в кабінетах електрокардіографії використовують подібний розчин, протираючи ним місце контакту електроду зі шкірою.

Концепція приладу реєстрації ЕКГ була вже зрозуміла і допрацьована на межі можливостей того часу.



Рисунок 1.8 – Електрокардіограф 1917 року

На рисунку 1.8 зображено прилад, за який Ейнтховен був назавжди записаний в історію медицини [4].

В 1932–1948 роках американський фізіолог Вільсон розробив методику однополюсних грудних відведень та їх векторний аналіз. На основі своїх досліджень він робить висновок про вигляд інфаркту міокарда на сигналі ЕКГ. З цього моменту починається відлік часу існування дійсно багатоканальної ЕКГ (3 основних, 3 однополюсних відносно середнього потенціалу і до 9 грудних відведень Вільсона, 3 з яких на спині).

Суть проста – вимір різниці потенціалів. Для цього на різні участки тіла людини накладаються електроди, а для кращої провідності на ділянках в місцях контакту наносять струмопровідний гель, який властивостями схожий на сольові розчини, які використовували в часи Ейнтховена та незадовго після нього. Електроди накладаються так, щоб вимірювати бажану різницю потенціалів – відведення. Вони є:

I – права рука (червоний електрод), ліва рука (жовтий електрод);

II – права рука, ліва нога (зелений електрод);

III – ліва рука, ліва нога;

aVR (права рука, ліва рука та ліва нога), aVL(ліва рука, права кука та права нога – чорний електрод), aVF(ліва нога, права та ліва рука) – підсилені однополюсні відведення, вимірюються відносно усередненого потенціалу трьох або двох електродів в залежності від системи, що використовується;

V1 – четверте міжребер'я справа від сонячного сплетіння;

V2 – четверте міжребер'я зліва від сонячного сплетіння;

V3 – посередині між V2 та V4;

V4 – п'яте міжребер'я по серединно – ключичній лінії;

V5 – на перетині горизонтального рівня 4-го відведення і передньої пахвової лінії;

V6 – на перетині горизонтального рівня 4-го відведення і середньої пахвової лінії;

V7 – на перетині горизонтального рівня 4-го відведення і задньої пахвової лінії;

V8 – на перетині горизонтального рівня 4-го відведення і серединно – лопаткової лінії;

V9 – на перетині горизонтального рівня 4-го відведення і паравертебральної лінії.

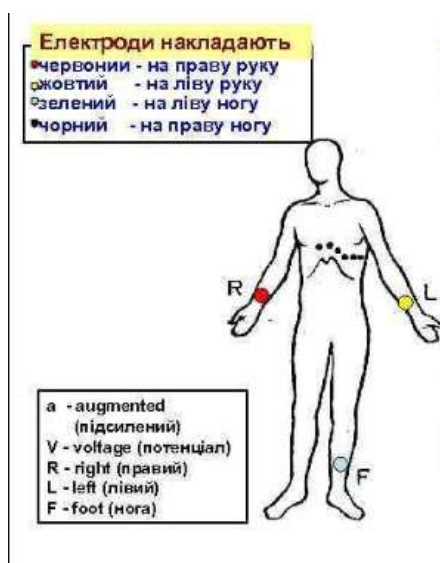


Рисунок 1.9 – Основні відведення ЕКГ на тілі людини

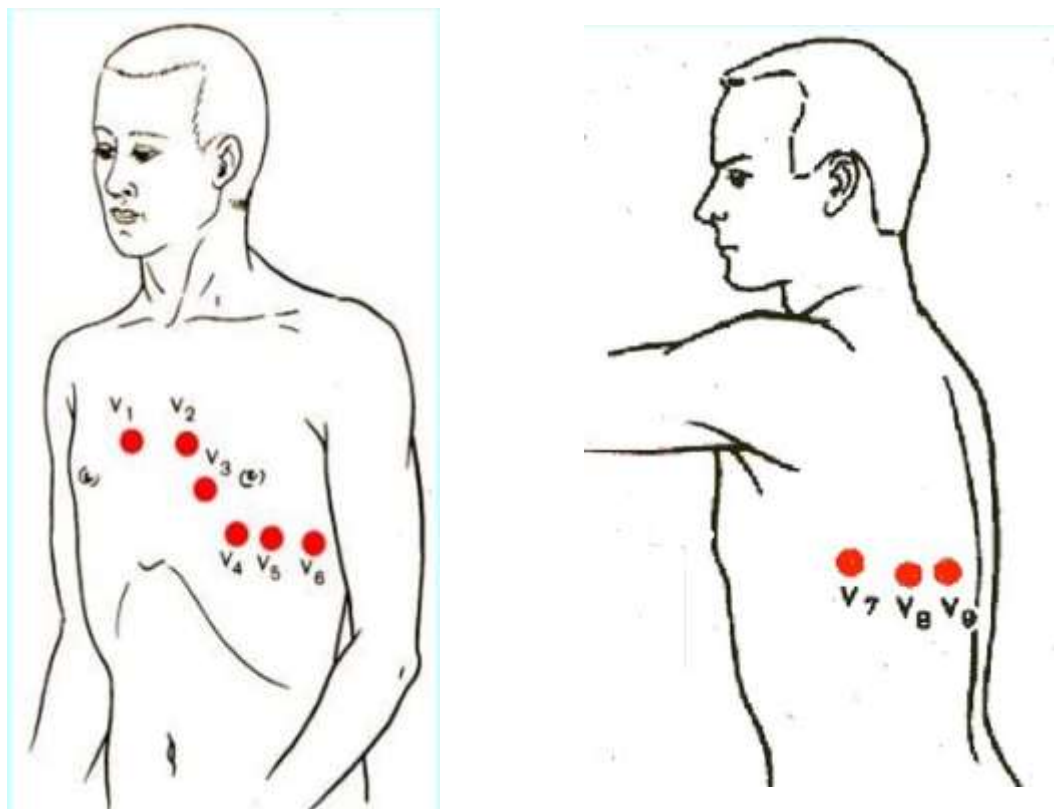


Рисунок 1.10 – Зображення грудних відведень на тілі людини

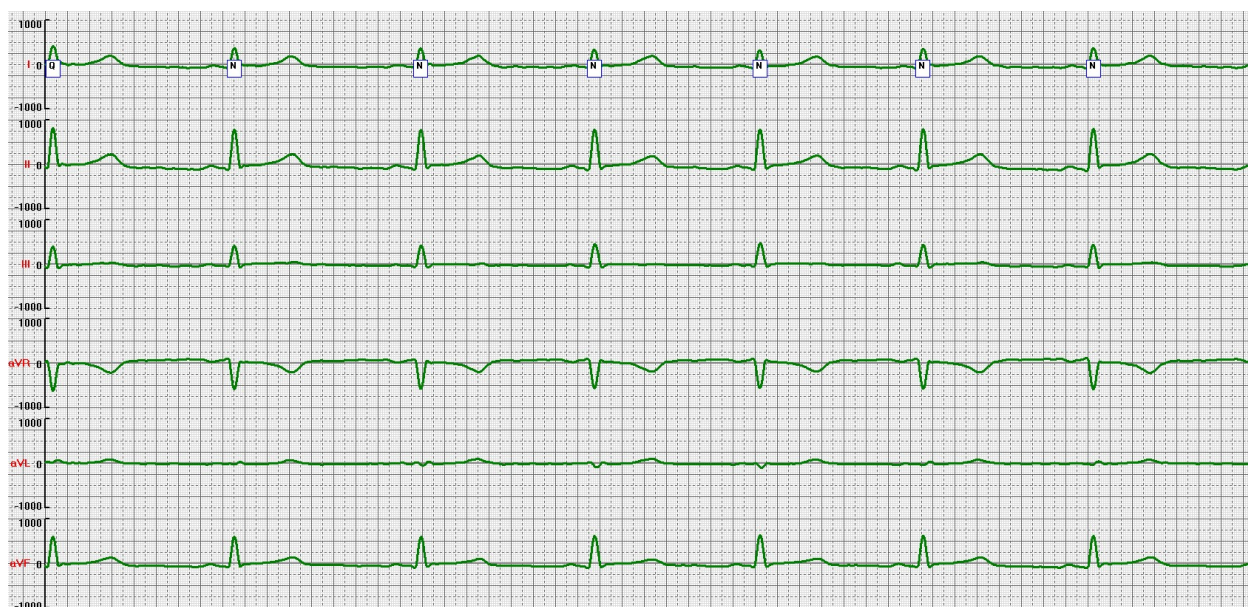


Рисунок 1.11 – Приклад 6-канального ЕКГ, використовуваного в роботі

Зараз на зміну самописців прийшли цифрові апарати. Замість паперу та чорнил використовують пам'ять приладів, записуючи туди масиви даних. Цей

спосіб зберігання та аналізу сигналів більш точний, адже зобразивши ці дані як на рисунку 1.11, їх можна масштабувати, вважати зображення векторним, що додає точності, адже растрове зображення має товщину лінії, що викликає деяку похибку спостереження, особливо на тонкій стрічці паперу, що іноді використовують і сьогодні.

Сучасні прилади ЕКГ не лише вимірюють сигнал, а й відразу його підсилюють та фільтрують. Використовуються смугові фільтри високих порядків для діапазону частот від 0.4 Гц до 40 Гц. Оскільки середня норма ЧСС 60 – 65 уд. с. в хв. (1 Гц) можна сказати, що смугу можна значно звужити, але на практиці користування фільтрами з вузькою полосою може спотворити сигнал. А діапазон, визначений вище, враховує і постійну складову сигналу, і шум 50 Гц, створюваний мережею, і високочастотні шуми сигналів. До того ж зробити такий фільтр з рівної «полицею» амплітудно частотної характеристики технічно простіше.

1.4 Приклад стандартного кардіографу

Розглянемо 12-ти канальний електрокардіограф UCARD 100.

Виробник – UTAS со. (ЮТАС), компанія заснована в Україні в 1992 році. Пройшла сертифікацію на відповідність стандарту EN ISO 13485:2016, а продукція сертифікована у відповідності з європейськими стандартами (Директива 93/42/ЕСС) та має маркування (CE 2409).

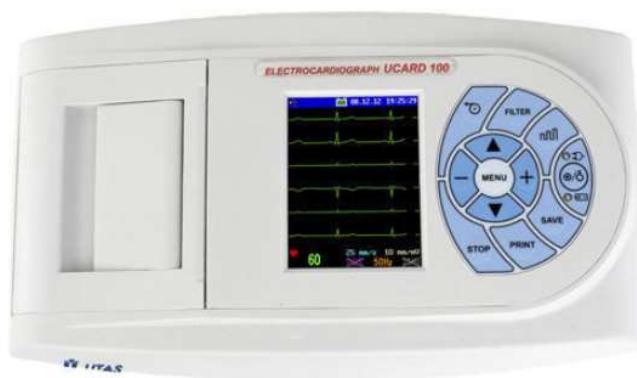


Рисунок 1.12 – Електрокардіограф UCARD 100

Наявна технологія UNET – система телеметричного комплексу, здатна виконати первинний аналіз та передати дані лікарю через спеціальні точки доступу і тут же отримати деякі рекомендації використовуючи зв'язок із телеметричним центром кардіології. Це дає можливість надати допомогу пацієнту за присутності будь-якого медичного персоналу, використовуючи вказівки віддаленого кардіолога.

Сигнал можна вивести на дисплей, передати на інші прилади або надрукувати на вбудованому швидкісному термо – принтері.

Наявні можливості накладання сигналів для порівняння сигналів, відслідковування динаміки змін, векторної електрокардіографії для отримання інформації наявності та характеру перенавантажень, гіпертрофії або дилатації передсердь.

Відносна компактність дозволяє використовувати апарат і в лікарнях, і в бригадах швидкої допомоги, до того ж наявний акумулятор, а вага всього 2 кілограми [5] (рисунок 1.12).

1.5 Приклади нестандартних методів отримання ЕКГ

Для прикладу візьмемо портативний електрокардіограф Cardiomo.



Рисунок 1.13 – Прототип приладу Cardiomo

Даний прилад здатний вимірювати артеріальний тиск, частоту серцевого ритму, температуру тіла, положення (орієнтацію в просторі) та рухливість пацієнта, а також одне відведення електрокардіограми.

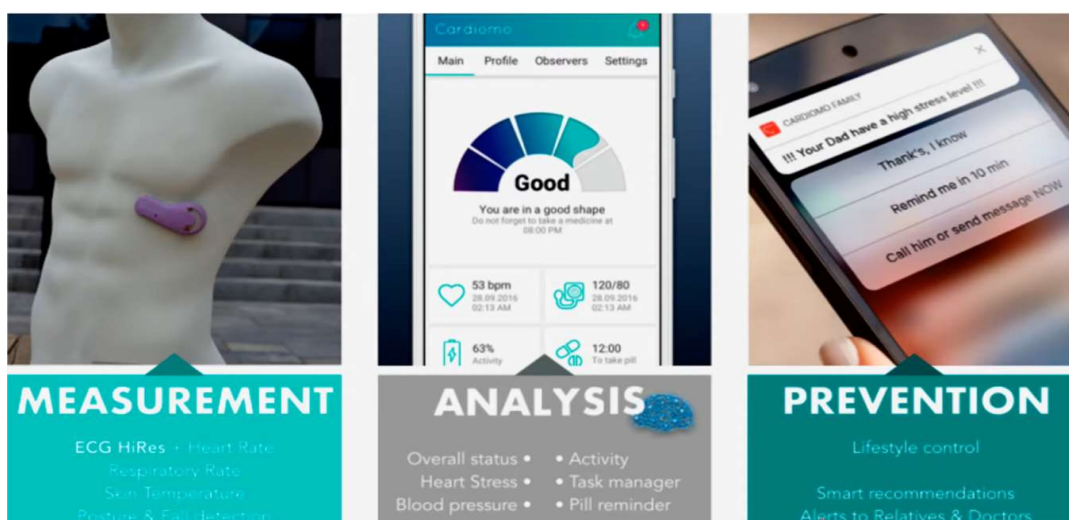


Рисунок 1.14 – Положення приладу на тілі людини, можливості в аналізі та функціоналі

Головною перевагою Cardiomo є максимальна компактність відносно інших портативних кардіографів. Як видно з рисунку 1.14, пристрій розташовується в зручному місці, де він не заважатиме пацієнту в повсякденному житті (розробники створили карту місць кріплення, які дозволяють знімати необхідне відведення). Утримується з допомогою «липких» електродів, які необхідно змінювати раз на півроку через забруднення поверхні. Після установки приладу алгоритм перевіряє якість сигналу, в разі поганого контакту чи сильно зашумленого сигналу додаток просить «переклеїти» пристрій. Кожні 10 хвилин Cardiomo фіксує різні дані вбудованими сенсорами та надсилає їх на телефон, тим самим збирає доволі широку статистику показників пацієнта. Тобто, даний прилад здатен цілодобово зчитувати інформацію про серцеву та дихальну діяльність людини. Дані передаються протоколами зв'язку на сервер, де і обробляються. В додатку власник приладу може відслідковувати такі параметри як варіабельність ритму серця, ЧСС, зміну температуру тіла, фази сну, активність та статистику протягом будь – якого періоду. До того ж можливо відслідковувати зміни в кардіограмі.

Важливою можливістю є спостереження за одним, необхідним лікарю відведенням. Однією з проблем ЕКГ є короткі сигнали. Використання подібного приладу може вирішити дану проблему, знімаючи довгі сигнали протягом цілого дня. В разі критичних відхилень фізіологічних показників додаток на телефоні запрошує підтвердження хорошого самопочуття. В разі негативної відповіді або її відсутності сповіщає про це лікаря, рідних або будь – яку людину вказану в додатку з допомогою СМС повідомлення [6].

Ціль даного приладу – швидкий та тривалий моніторинг стану пацієнта. Вимірювання 12-ти канальної ЕКГ не є технічним завданням інженерів даного винаходу. Звичайно, шумів більше, відведення лише одне, але виробник стверджує, що точність є достатньою для того, щоб приймати ці дані в тривалих дослідженнях функціонального стану лікарем, оскільки на даний момент розроблено доволі потужний математичний апарат для автоматичного аналізу отримуваних сигналів [7].

Якщо даний прилад здається достатньо компактним, то які перспективи відкриті перед приладом, зображеним на рисунку 1.15.

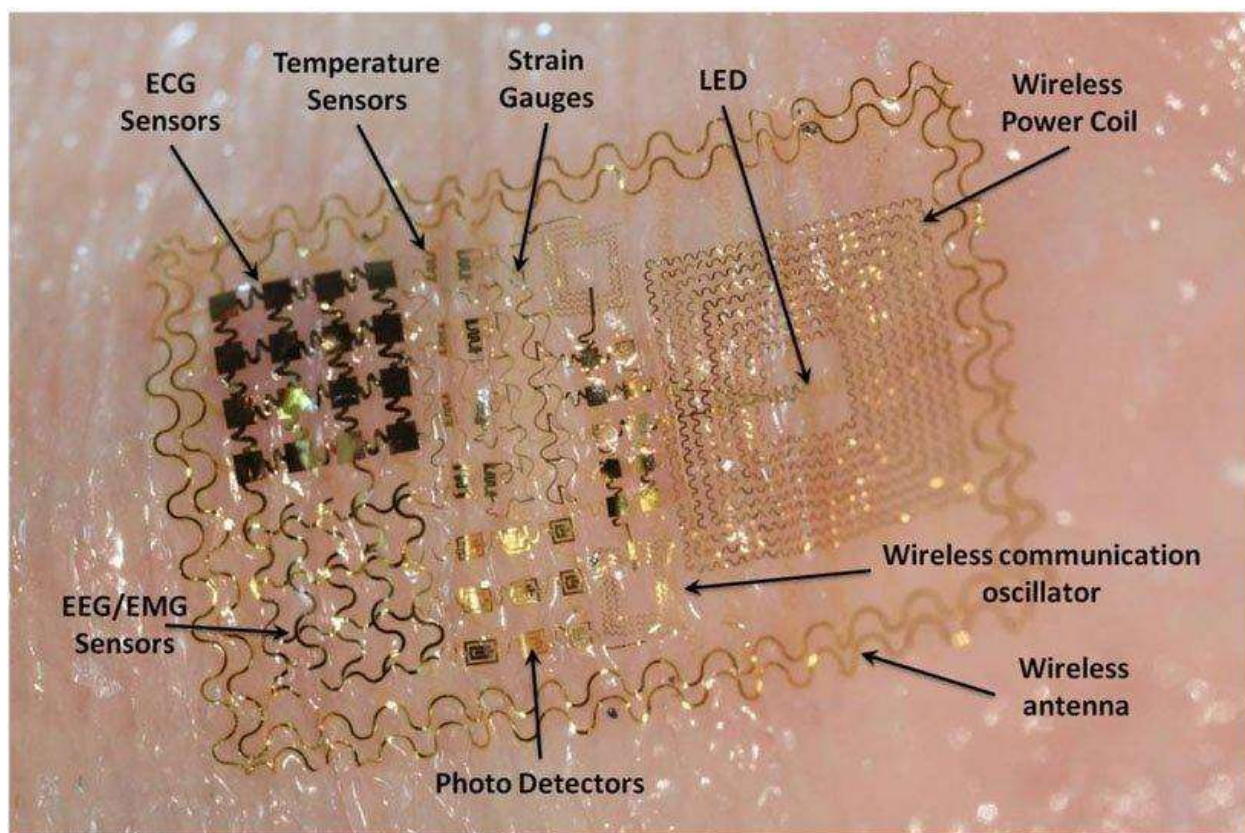


Рисунок 1.15 – Медичне електронне татуювання

На рисунку 1.15 наведене електронне татуювання. Це плівковий прилад, який наноситься на шкіру людини та здатен виконувати деякі функції. Даний прототип може вимірювати температуру тіла, знімати активність серця або будь-якого іншого м'язу, знімати сигнали електроенцефалограми (ЕЕГ) та електроміограми (ЕМГ), випромінювати світло та поглинати світло. Звичайно, ці всі функції не можуть використовуватись одночасно. Щоб відслідковувати активність серця необхідно помістити прилад в зону грудної клітки, для зняття ЕЕГ – на голову. Прилад може знаходитись на шкірі протягом 3-10 днів, поки шар шкіри під «татуюванням» не відпаде природним чином.

Отримані значення передаються з допомогою вбудованого осцилятора, який створює коливання. Далі необхідно вловити ці коливання та перетворити в дані, які цікавлять дослідника.

Прилад виконується з ниток міді, вуглецевих нанотрубок, фулеренів, графену та кремнієвих плівок.

Живлення даної схеми можливе в декількох варіаціях:

1. Фотоелементи генерують носії від світла, яке потрапляє на їх поверхню;
2. Живлення індукованим струмом з «катушки живлення» (як в RFID картках);
3. Перетворення різниці потенціалів на шкірі в потужність для живлення.

Останній метод знаходиться в фазі активної розробки. Адже подібне живлення можна використовувати, наприклад, в кардіостимуляторах замість радіоактивних елементів.

Найбільша проблема даного приладу – наднизька потужність. Тобто, немає сенсу зменшувати розміри вимірювального приладу, якщо надалі необхідно використовувати надточні системи, щоб зчитувати дані в реальному часі, оскільки подібне рішення не може мати пам'яті для зберігання даних.

1.6 Усереднений сигнал ЕКГ

Signal – averaged ECG (SAECG) – усереднений сигнал ЕКГ. Існує декілька варіантів його отримання, але всі вони базуються на одному принципі: зі стандартної ЕКГ обирається ділянка, яка виділяється в кожному QRS комплексі, а потім знаходиться середнє значення кожної точки. Тобто можна взяти інтервал R-R та знаходити усереднений сигнал серед цілого комплексу, що використовується найчастіше, можна взяти лише зубці QRS та «нарізати» інтервали по ним, так само можливо зробити і з допомогою будь-якого яскраво вираженого інтервалу. В даній

роботі буде використані сигнали, отримані саме таким методом – виділення R-R інтервалу та їх усереднення.

Ще одним методом отримання усередненого QRS є вирізнання схожих комплексів з «ідеальним». Ідеальний тут – перший чіткий комплекс, який є в сигналі ЕКГ, який береться за основу для алгоритму, який порівнює всі наступні та відбирає максимально на нього схожі. Ідеал може бути вибраний лікарем або автоматично. Якщо відсоток подібних комплексів серед усіх великий, їх усереднюють та приймають за середній. Такий підхід відразу може вказати на сильні відхилення симетрії ЕКГ. Тобто, якщо схожих комплексів мало, то проблемою пацієнта може бути критичною і пов'язаною з сильними патологіями, наприклад фібриляція передсердь. Хоча даний метод дозволяє оцінити даний параметр, на практиці він може не нести цінності, оскільки настільки сильні відхилення відразу бачить лікар, або вже знає про них. В такому випадку не доцільно використовувати аналіз усередненого комплексу.

Аналіз SAECG частіше всього використовується у випадках, коли необхідно розглядати малоамплітудні коливання, які при розгляданні в кожному окремому комплексі можуть бути сильно зашумлені, навіть після фільтрації. Адже однією з головних переваг SAECG є позбавлення сигналів від шуму за рахунок виділення найбільш вагових амплітуд.

Однією з перших згадок про використання SAECG було дослідження [9], опубліковане в 1973 році.

Головною причиною використання даного методу дослідження – зменшення шумів в усередненому сигналі. Тобто кардіокомплекс, який був середнім серед усіх представлених на записі фільтрував шуми сам. Шум має випадкову природу, а тому має маленьку вагу в обрахунках усередненого комплексу. Далі достатньо застосувати, наприклад, фільтр Баттерворта зі смугою пропускання від 0.7 до 5 Гц (на практиці полоса ширша для простішого досягнення рівної характеристики та відсутності можливості спотворити сигнал) для позбавлення мережевого шуму.

Автори зазначають, що цей метод має одну недооцінену властивість – фільтрація, так званого, психологічного шуму. Цей шум викликаний

переживаннями пацієнта. Людина перестає хвилюватися при вигляді медичного обладнання лише тоді, коли починає працювати з ним щоденно. Будь-який похід до лікаря супроводжується підвищеним серцебиттям, зміною артеріального тиску і тремором м'язів, який неодмінно може стати джерелом шумів в підключеному електроді.

Друга особливість – можливість дослідження стану реполяризації передсердь. Вона супроводжується малоамплітудними коливаннями які практично непомітно на одиничному циклі серцевої активності. Сигнали повторюються та іноді змінюють свою тривалість. За рахунок цього даний сигнал відображається як малоамплітудний шум, але тільки на одному відрізку комплексу.

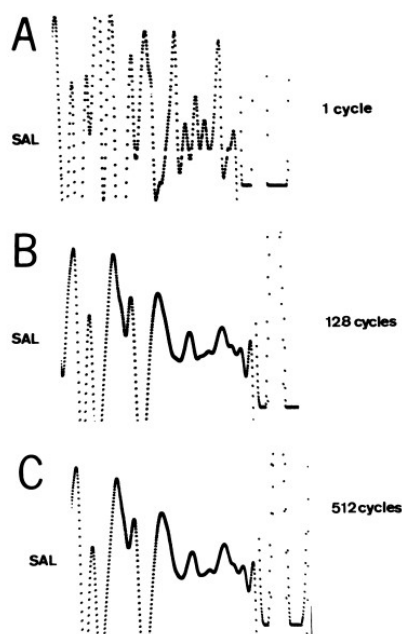
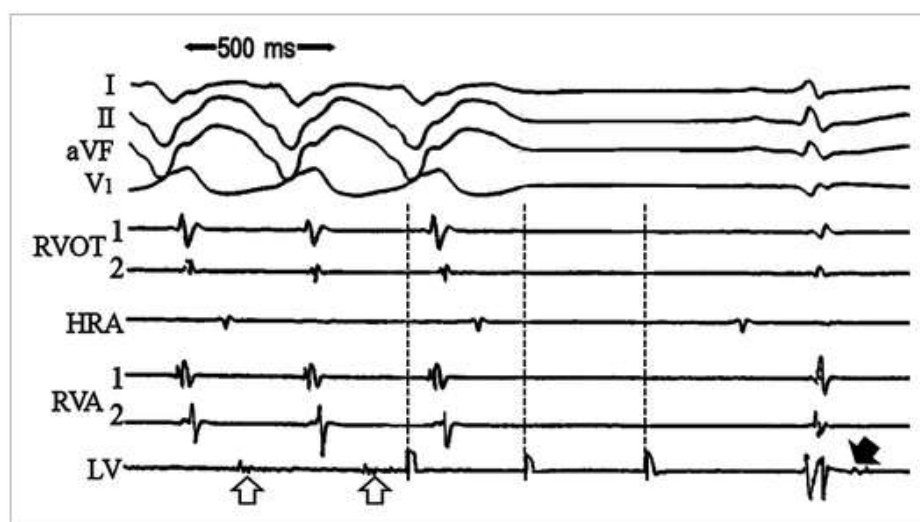


Рисунок 1.16 – Кількість накладених сигналів та відповідний результат усереднення

З рисунку 1.16 видно, що із зростанням кількості ітерацій деякі сегменти набувають більш чіткого вигляду. Це говорить про те, що в більшості інтервалів вони співпадають, а ті, що менш чіткі – ні. Це дає можливість досліднику використовувати статистично найбільш стійкі показники. Це підтверджується фактом, що у досліді, наведеному на рисунку після 128 ітерацій відношення сигнал-шум збільшилось в 11 разів. Після 512 ітерацій ще в 2 рази (рисунок 1.16).

Усереднені кардіограми почали використовувати при дослідженні ішемічної хвороби та прогнозування і подальшого упередження інфаркту міокарда [10].

Суть полягає у математичному об'єднанні трьох ортогональних відведень в одне, яке в результаті покаже малоамплітудний сигнал після основного комплексу. Амплітуда даного сигналу складає менше ніж 40 мкВ і є наявним у відфільтрованому сигналі, а значить має постійну частотну характеристику, подібну активності серцевого м'язу.



Стрілками відмічені слабкі сигнали, які отримуються в результаті об'єднання трьох відведень.

Рисунок 1.17 – Усереднений кардіокомплекс

На рисунку 1.17 наявні сигнали ЕКГ пацієнта зі старим інфарктом міокарда передньої стінки. Спочатку ми бачимо основні відведення, а потім ендокардіальні електрокардіограми правого шлуночкового відтоку (RVOT), правого передсердя (HRA), верхівки правого шлуночка (RVA) та ураженого інфарктом лівого шлуночка (LV). Під час синусоїдального ритму є пізні фрагментовані низькоамплітудні сигнали, описані раніше. Маючи вентрикулярну тахікардію фрагментовані потенціали зони лівого шлуночка продовжують свою тривалість, нав'язуючи діастолічну фазу серцевого циклу і зникають перед наступним збудженням шлуночка.

Варто зазначити, що тахікардія – не хвороба, а стан, коли частота серцевих скорочень стає більшою за 90 ударів серця за хвилину. Це не завжди вказує на погану роботу серця. Важливо розуміти, що фізичні навантаження, переживання чи вплив стимуляторів викликають тахікардію. Важливо одне – «синхронна» тахікардія шлуночків та передсердь.

До цього малопотужного сигналу ще в 1994 році були запропоновані деякі критерії, які дозволяють виявити даний сигнал, вони включають: відфільтрований $QRS \geq 114$ мс, $RMS-40 < 20$ мкВ (Rated Maximum Sinusoidal)) та $LAS-40 > 38$ мсек, при цьому специфіка методу залежить від кількості позитивних критеріїв. Важливо, щоб виміряні сигнали обов'язково мали вузькі комплекси в кардіограмі з 12 відведень. Оскільки дані критерії дуже строгі, вони не завжди виконувались у пацієнтів, тому що деякі пацієнти мали дуже широкі комплекси зі зменшеною амплітудою та неспецифічними дефектами внутрішньошлуночкової провідності пізніше були введені модифіковані критерії $QRS \geq 145$ мс, $RMS-40 < 17,5$ мкВ та $LAS-40 > 50$ мс.

Зміни цього сигналу можуть стати попередженням про майбутню тахіаритмію шлуночків, що може призвести до порушень циклу та клінічної смерті. Найбільш гостро це питання стоїть у людей, які пережили інфаркт і мають максимально попередити можливість рецидиву.

Значення SAECG було додатково підтверджено аналізом в статистиці групи пацієнтів з тахікардією, які пережили інфаркт міокарда. Ненормальна SAECG виявляє підвищений ризик для набору аритмії та повторного інфаркту. Що, на жаль, підтверджується в 44% випадків на дистанції 5 років.

На сьогоднішній день дослідження усередненої електрокардіограми має важливе місце у попередженні ішемічної хвороби. Воно включене в стратегії зниження ризику, що ще більше підкреслюється нещодавним дослідженням [11].

Як зазначено вище, на сьогоднішній день аналіз усередненого комплексу використовується для дослідження стану людей, які перенесли інфаркт міокарда. Адже їх функціональний стан практично повністю відображається на низькоамплітудних коливаннях, які найкраще можна відстежити саме на

усередненому комплексі ЕКГ. Автори роботи [12] запропонували використовувати дещо інші фільтри під час підготовки сигналів. В результаті були отримані більш чіткі зображення остаточних коливань, які і необхідні для подальших дослідів. Маючи ці сигнали лікар може аналізувати їх динаміку з більшою точністю. Запропоновані фільтри: фільтр верхніх частот з частотою зрізу $f_c = 25$ Гц. При порівнянні виявилось, що дана конфігурація виявила остаточні коливання в більшій кількості ЕКГ, ніж 40 Гц. Враховуючи те, що використані дані були виключно сигналами пацієнтів, що перенесли інфаркт міокарда можна сказати, що даний фільтр більш якісний, але, як зазначають автори статі «лише в дослідженні подібних відхилень».

Dilated cardiomyopathy (DCM) характеризується дифузним фіброзом, протиставляючи зазвичай добре відмежованому субстрату, що оточує інфарктний шрам; отже, складні шлуночкові тахіаритмії можуть виникати з декількома механізмами, з різною частотою захворюваності, що спостерігається в різних групах DCM. Внаслідок цієї складної патофізіології стратифікація ризику залишається проблемою у цих пацієнтів, і кілька проблем викликають значення аритмічних маркерів у DCM; насправді багато центрів часто опускають електрофізіологічну оцінку в DCM, вибираючи використання LVEF як єдиного критерію. Оскільки декілька досліджень показали не надійність даного методу на великих проміжках часу (20 – 30 років) та пацієнтів, які не мали інфаркту. Що стосується застосування для пацієнтів з інфарктом міокарда, то група неінвазивних показників може керувати запрограмованою шлуночковою стимуляцією. Даний підхід потребує подальшого дослідження.

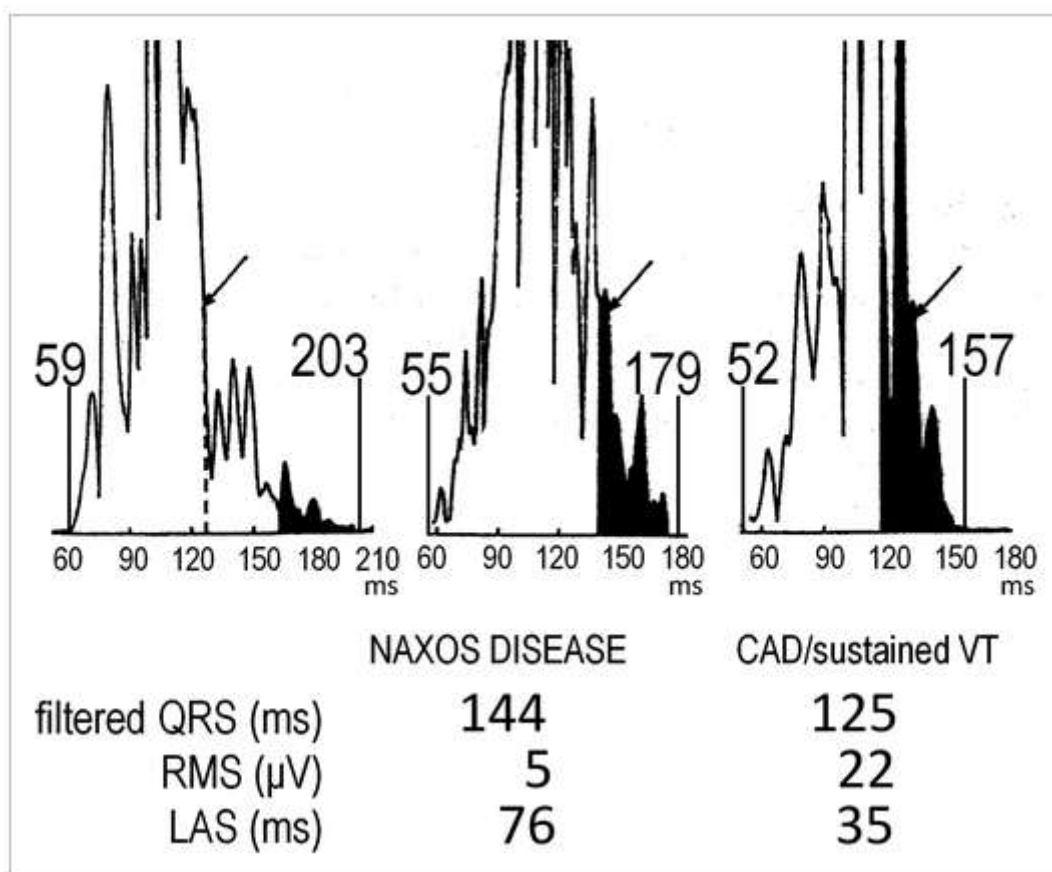


Рисунок 1.15 – SAEKG пацієнта з Наксосовою хворобою (зліва), після інфаркту міокарда (посередині) та клінічно здорового

Тут явно видно різницю між здоровою людиною, та тою, що пережила інфаркт. Але є ще одна важлива деталь. Присутні навіть явні розходження між здоровою людиною та пацієнтом з шкіряним захворюванням. А це означає, що SAEKG може вловлювати вплив багатьох факторів. Цю тему почали розвивати і прийшли до того, що даний показник може бути вирішальним у визначенні тренуваності серця та подальшого потенціалу. Це детальніше описується в роботах [13] та [14].

1.7 Постановка задачі роботи

З огляду на опрацьовану інформацію можна зробити висновок – аналіз ЕКГ, в більшій мірі, необхідно використовувати для визначення наявності та виду патологій у пацієнта. На жаль, аналіз функціонування серця здорової людини не так розповсюджений, як опрацювання пацієнтів з проблемами. Тому необхідно розробити методи аналізу, які покажуть на скільки здорова серцево-судинна система, на що варто звернути увагу та яка ймовірність того, що скоро необхідно буде звернутися до фахівців за лікування яких небудь відхилень від норми. До того ж, необхідно визначити і саму норму. Звичайно це вже було зроблено ще фахівцями кардіології 70-100 років тому, але останнім часом, як зазначають сучасні фахівці, існує зміна цієї норми, відносно тієї, що зазначена в наукових джерелах, які складають фундамент знань нових фахівців. Норма не є сталою, серце, як і весь людський організм пристосовується до навколишнього середовища та навантаження. Не є секретом і те, що еволюція не зупинилась, зміни людства тривають досі і будуть тривати довго після нашої ери. Враховуючи вищесказане, можемо припустити, що варто спробувати виконувати аналіз сигналів дещо іншими методами: відійти від рутинного перегляду і побудувати систему, яка буде роздивлятись дані сигнали з іншого боку, з боку «машинного сприйняття». Враховуючи рівень абстракції алгоритмів сучасних нейронних мереж, важко навіть оцінити що вони бачать в цих сигналах та яку вагу приділяють кожному параметру, кореляцію між якими значеннями проводять і т.д.. Тобто використати методи машинного навчання для визначення норми серед людей сьогодення. Визначити величину «здоров'я серця» в порівнянні з іншими людьми, які вважаються клінічно здоровими в результаті медичного обстеження.

1.8 Висновки до розділу 1

Було розглянуто методи та техніку електрокардіографії, її використання, яким чином можна отримати сигнали та потреби в її аналізі. ЕКГ хворого пацієнта необхідно досліджувати, щоб знайти причину відхилень та спробувати їх подолати, або попередити подальший негативний розвиток. ЕКГ здорової людини важливо аналізувати для прийняття деяких превентивних заходів з ціллю попередити перехід пацієнта до стану хворого. Було визначено що таке SAECG, для чого її намагались використовувати колись і для чого вона потрібна сьогодні. На основі отриманих знань спробуємо визначити поняття норми та відповісти на питання здорових людей : «На скільки я здоровий?». Розв'язувати дану проблему будемо за допомогою автоматичного аналізу методами машинного навчання. В наступному розділі розглянемо існуючі методи та виберемо той, що підходить для розв'язку поставленого завдання в пункті 1.7.

2 МЕТОДИ МАШИННОГО НАВЧАННЯ В АНАЛІЗІ SAECG

2.1 Штучний інтелект, машинне навчання, методи та алгоритми машинного навчання

Поняття штучного інтелекту досить розмите. Більша частина сучасної техніки має мікročіп, а виробник говорить про наявність штучного інтелекту в своїх приладах. Але в більшості це лише вибіркове і обмежене копіювання можливостей людини, її поведінки на штучно створеному об'єкті для спрощення роботи самої людини, що не є проявом машинного інтелекту. До цих приладів можна віднести роботизовані руки (маніпулятори), конвеєри, елементи «розумного дому», тощо. Інтелект походить від латинського поняття *intellectus* – розум, розсуд, здатність мислити. Штучний інтелект (Artificial Intelligence – AI) розуміється як здатність автоматичних систем брати на себе функції людини, приймати оптимальні рішення на основі раніше отриманого досвіду і аналізу зовнішніх чинників. Мислення – це процес перетворення та обробки інформації, емоцій, тощо. Інтелект і мислення пов'язані багатьма цілями та задачами: розпізнавання ситуацій, логічний аналіз, планування поведінки. Характерною є наявність здатності до навчання, накопиченню досвіду, адаптація до умов, що змінюються. З цього визначення випливає основна проблема штучного інтелекту – можливість чи неможливість моделювання мислення дорослої людини чи дитини [15]. Тобто штучний інтелект повинен мати поведінку біологічного інтелекту, для цього необхідно створити математичну модель, яка описує процеси в системі організму, що сприймає, обробляє, передає і зберігає інформацію, тобто створити модель нервової системи. Умовна одиниця нервової системи – нейрон. Вони утворюють біологічну мережу, а математична модель кожного нейрону та їх взаємодії називається штучною нейронною мережею.

Перші дослідження в даній області були зроблені відразу після появи перших електронно обчислювальних машин. Період народження науки про штучний інтелект припав на 1943 – 1956 роки, цьому передували відкриття в нейрології, кібернетиці, теорії сигналів та теорії алгоритмів Алана Тюрінга. Тобто було відомо, що мозок – мережа нейронів, що передає інформацію по принципу «все або нічого», 0 або 1, Норберт Вінер описав цифрові сигнали, Тюрінг запевняв, що будь-які обрахунки можуть виконуватися з допомогою цифрових операцій. Уолтер Пітс і Уоррен МакКаллок проаналізували мережі, які складаються з ідеалізованих штучних нейронів, вони показали як мережа може виконувати прості логічні функції. Натхненний цим успіхом Марвін Мінські в 1951 році будує першу нейронну мережу з випадковим з'єднанням. В цей час Френк Розенблат описав модель людського сприйняття і назвав її «перцептроном». Нейрофізіолог заявив, що будь-яка інформація, що поступає до мозку асоціативним методом порівнюється з накопиченою, після чого мозок виносить рішення у вигляді реакції, яку вважає найбільш вірною [16]. Це і стало основою створення моделі (рисунок 2.1).

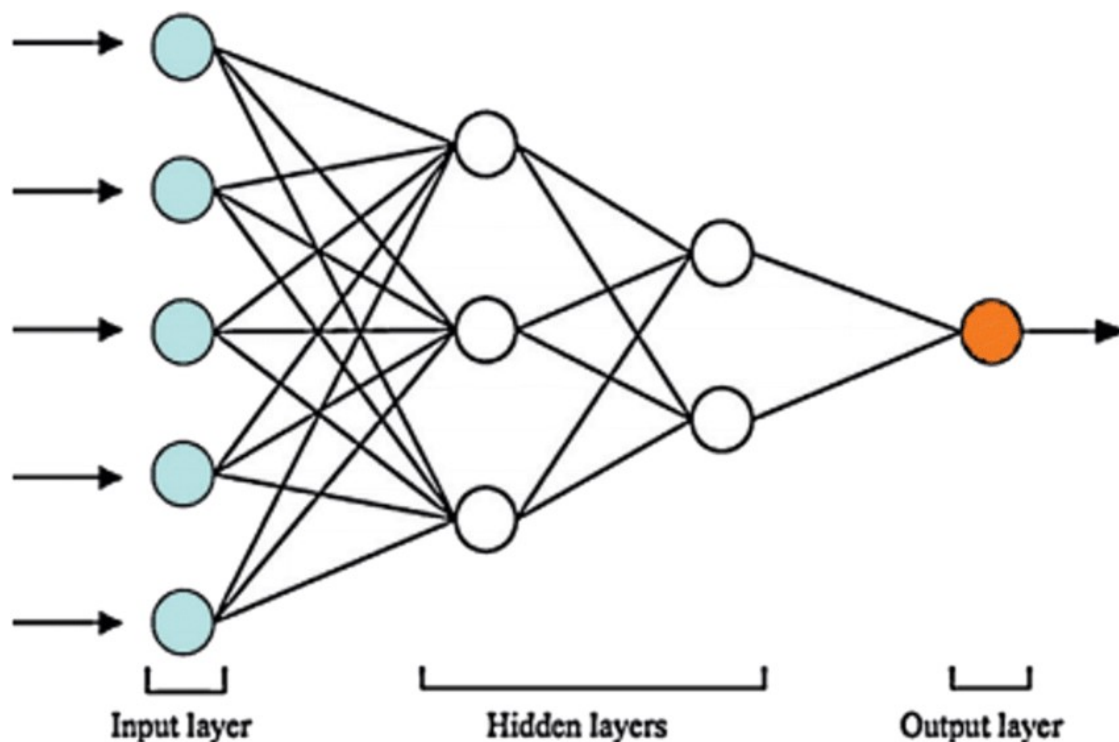


Рисунок 2.1 – Модель перцептрона

З рисунку 2.1 видно, що інформація поступає на вхідний шар (наприклад тактильні рецептори на кінчиках пальців), далі – до прихованих асоціативних шарів, що накладають на дані певну вагу, яка визначається з досвіду (подібних ситуацій, що зустрічались раніше), а потім отримані значення сумуються і поступають до останнього шару – функції активації, від якої залежить вихідна реакція на сигнал (прибрати руку, відчуваючи слизьку поверхню, асоціюючи її з покривом безхребетних або подібних речей, що викликають страх, огиду і т.д., але асоціації можуть бути різними, від них і залежить остаточне рішення).

Порівняння з прецедентами не є єдиним механізмом нейронних мереж як біологічних, так і штучних. Існує метод прогнозування результату вибору, тобто обираючи рішення ситуації система моделює подальші дії та їх наслідки. Наприклад, граючи в шахи хороший гравець аналізує ризики кожного свого можливого ходу і прогнозує хід суперника виконуючи аналіз ситуації, яку він моделює для себе, вважаючи прийнятим попереднє рішення. На такому принципі була побудована мережа Deep Blue. Алгоритм прораховував можливі рішення спираючись лише на ризики втрати цінності або створення програшної для себе ситуації використовуючи тільки правила гри та призначення ваги кожній своїй фігурі на дошці в даній та можливих ситуаціях після свого ходу. Алгоритм догравав тисячі варіантів закінчення партії та обирав найкращий для себе. Таким чином в 1996 році Deep Blue програла «битву машини і людини» проти Гарі Каспарова, який був кращим в шахах протягом 13,5 років на планеті Земля, (рахунок 2:4), а через рік взяла реванш ($2\frac{1}{2}:3\frac{1}{2}$). Швидкість прорахунку складала приблизно 200 млн позицій в секунду, що в складній ситуації затримувало хід на 5 хвилин, хоча були моменти очікування в 14 хвилин. Трохи пізніше був створений алгоритм DeepMind, його принцип побудований на аналізі тисяч ігор професійних гравців, відносно яких і обирається наступний хід. В даному алгоритмі використана система машинного навчання, що створює деякий досвід мережі. Дана система характеризується не прямим рішенням задачі, а навчанням, в процесі якого використовуються приклади вирішення подібних задач.

Методи машинного навчання будуються на різноманітних чисельних методах, теорії ймовірності, теорії графів, математичної статистики, аналізі даних в цифровій формі і т.д.. Ключовою особливістю даного класу штучного інтелекту є досвід, який необхідно забезпечити. Для цього були створені алгоритми навчання, вони можуть бути двох типів:

1. Детермінований – ітераційно коректує параметри мережі, які засновані на її поточних параметрах, вхідних величинах, фактичних та бажаних виходів. Яскравим прикладом є метод зворотного розповсюдження похибки.
2. Стохастичний – змінює параметри мережі випадковим чином. При цьому зберігаються лише ті зміни, які призвели до покращення результатів.

Способів машинного навчання є велика кількість: навчання з вчителем, навчання без вчителя, навчання з підкріпленням, активне навчання, навчання з частковим втручанням вчителя, багатозадачне навчання, багатоваріантне навчання, транспродуктивне навчання.

Вчителем виступають тренувальні дані, де на кожні початкові умови, описуючі завдання є відомий розв'язок. Зазначимо, що для формування необхідного відображення розмір вхідних даних та відповідних відповідей має бути достатньо великим. Навчання з підкріпленням також передбачає наявність прийнятого рішення. Активне навчання дозволяє алгоритму самостійно обирати ситуацію, на яку отримає вірне рішення. Метод транспродуктивного навчання та з частковим втручанням вчителя особливі тим, що тренувальна вибірка частково має в наявності вірні відповіді, а частково – ні. Метод навчання без вчителя для задачі даної роботи більш правдоподібний, адже важко уявити, що в реальному мозку є деякі механізми, що порівнюють отримані результати з бажаними. Алгоритми навчання без вчителя застосовуються коли є лише вхідні дані – ситуації, а розв'язок не відомий. На їх основі мережа вчиться давати найкращі значення виходів. «Найкраще значення» визначаються алгоритмом навчання, параметри якого підстроюються так, щоб мережа видавала однакові результати для достатньо близьких вхідних значень [17].

Методи машинного навчання використовуються для вирішення задач розпізнавання образів (візуальних, звукових, цифрових), прогнозування ситуацій (регресія), кластеризації, моделювання, виявлення аномалій (outlier detection), зменшення розмірності. Застосування нейронних мереж можливе в програмуванні, наукових дослідженнях, машинобудуванні, безпеці людини та майна, виробництві, розвагах, побутових приладах та медицині. Аналіз слів в тексті, моделювання існування популяції при різних параметрах виду, підказки можливих відповідей оператору колл центру відносно запиту людини, що дзвонить, виявлення наявності kota на зображенні, підбір відео на відомому відеосховищі YouTube і багато де ще.

Для медичної сфери були створені системи виконання точних операцій і консультації лікаря в складних ситуаціях. Сьогодні суспільство цікавиться системами прийняття рішень в реальному час, засобами зберігання, вилучення, аналізу та моделювання знань, системами динамічного планування. Серед них уже є конкретні результати:

1. DENDRAL – високоінтелектуальна система розпізнавання хімічних структур. Найстаріша з експертних програм, перші версії з'явилися в 1965 році. Подаючи на вхід інформацію про речовину, дані спектроскопії користувач на виході отримує вигляд відповідної хімічної структури[18].

2. MICIN – експертна система медичної діагностики. Програма ставить діагноз та рекомендує курс медикаментозного лікування відносно представлених симптомів[19].

3. PUFF – система аналізу порушень дихання людини. Вона базується на системі MICIN[20].

На сьогоднішній день найкраще зарекомендували себе мережі з сильно обмеженою областю приймання рішень. Для моделювання повноцінного інтелекту, здатний на самостійне існування, який в результаті може стати основою побудови нового розумного виду необхідно побудувати надскладну систему, яка здатна наділяти сенсом окремі дані, розуміти емоції, психічний стан, поняття волі, тобто робити все те, що притаманно лише живому організму. У алгоритмів немає

бажання самовдосконалюватись, вони мають задачу і виконують її з максимально якісним для людини результатом.

2.2 Застосування методів машинного навчання в аналізі ЕКГ

В аналізі ЕКГ з допомогою машинного навчання можна вирішувати завдання виявлення аномалій (пошук хворого серед здорових), виділення окремих патологій, окремих параметрів, швидкий аналіз, моделювання, прогнозування, симуляції [21].

Зараз все частіше використовують електронний метод зберігання сигналів, це означає, що кожен наступний пацієнт, якому вимірюють сигнал активності серця потенційно може поповнити базу даних для тренування алгоритмів машинного навчання. У задачах пошуку параметрів в кожному окремому сигналі проблеми з форматом вхідних даних немає. Завдання кластеризації та пошуку аномалій накладають деякі обмеження на вхідні дані. Для цього можна використовувати сам сигнал з усіма його відведеннями, але в даному підході є великий недолік – необхідність перерахунку сигналів у масиви значень однакової довжини, що може сильно спотворити сигнал, як наслідок – втрата коректних часових показників (тривалість P, P – R, QRS, R, S, Q, TR , UT). Інший механізм – використання характерних ознак. Це амплітудні показники (амплітуда P, Q, R, S, T, U, залишкових імпульсів), часові та інші (площа під P, QRS комплексом і тому подібні). Таким чином забезпечується фіксована кількість вхідних даних. Самі ознаки можна знаходити вручну, методами машинного навчання, або методами пошуку локальних точок із заданими параметрами.

Алгоритми машинного навчання, які використовуються для ЕКГ – це потужні інструменти для надання допомоги в задачах оцінки стану пацієнта та мають потенціал для швидкої стратифікації пацієнтів, використовуючи переваги накопиченої інформації. Наприклад визначення типу аритмії з допомогою алгоритмів штучного інтелекту. Дослідження [22] описує результати використання

алгоритмів кластеризації для визначення виду порушень ритму серед багатьох наявних у вхідних даних.

Використовувана база даних (MIT-BIH) налічує 92 пацієнти (40 чоловіків та 52 жінки, середній вік $39,75 \pm 19,06$ років). В наявності наступні порушення ритму серця : синусоїдальна брадикардія, вентрикулярна тахікардія, синусоїдальна аритмія, передчасне скорочення передсердь, темп биття, фібриляція та трепетіння передсердь. Використовується лише R-R інтервал.

Автори пропонують використовувати FCM clustering in fuzzy self-organizing перед входом. Ідеєю є об'єднання самоорганізуючого шару з MLP(multi-layer perceptron) для збільшення точності класифікації. В результаті досліджень було виявлено, що оптимальна кількість прихованих вузлів – 48. Ця кількість забезпечує точність класифікації в 99.9%. Вона була визначена експериментально, порівнюючи результати тренувань з 67, 106 та 212 сегментів .

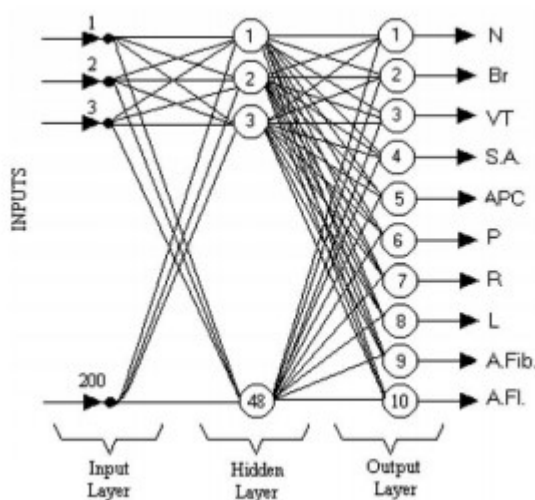


Рисунок 2.2 – Оптимальна архітектура нейронної мережі

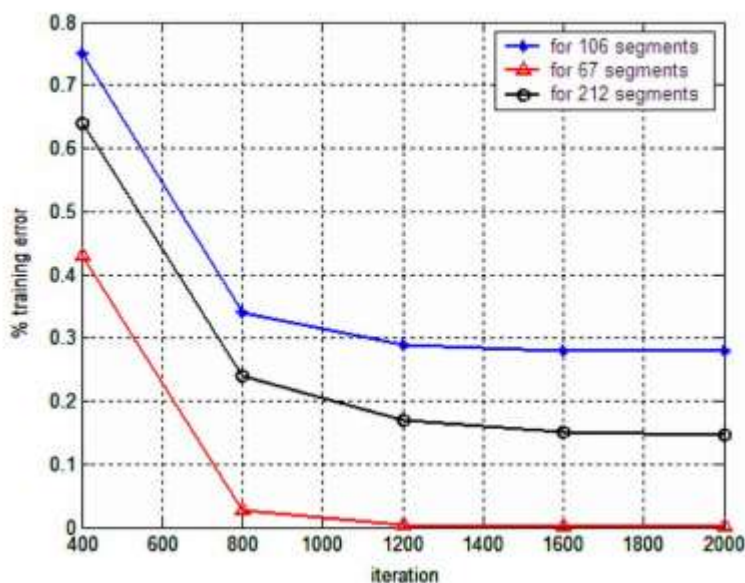


Рисунок 2.3 – Результати класифікацій з трьома різними налаштуваннями мережі

Як видно з рисунку 2.3, більша кількість сегментів призводить до наближення похибки в деякому значенні, що не дозволяє отримати точний результат.

Як висновок було порівняно дві архітектури. FCNN(fuzzy clustering neural network) та NN(neural network), тобто перцептрон з додатковим вхідним шаром та без. Додатковий шар збільшив точність класифікації на 1%. Тобто, за 2000 ітерацій нейронна мережа має точність не 98.9%, а 99.9%. З одного боку результат не значний, всього лише один відсоток, в порівнянні з раніше запропонованою архітектурою. Але враховуючи, що це питання в області медицини, а отже і здоров'я людини, де помилка в один відсоток може стати невідвратною точкою однієї з сотні людей.

2.3 Вибір методу для використання в даній роботі

Для розв'язку завдання даної роботи необхідно обрати метод який здатен дати бажаний результат виходячи з наявних початкових даних: усереднені кардіокомплекси для шести відведень, набір з 34 параметрів, розрахованих з комплексів.

Спочатку виберемо тип завдання, яке має розв'язувати нейронна мережа. Як описувалось раніше, необхідно створити алгоритм, який визначає «норму здоров'я» та вказує на скільки пацієнт далеко від цієї норми. Це можна виконати трьома методами:

1. Класифікація;
2. Кластеризація;
3. Пошук аномалій.

Класифікація – розподіл по групах за спільними ознаками. Класи можуть поділятися на підкласи, види, тощо. Можна провести класифікацію пацієнтів, за ознаками притаманними здоровим та хворим. Виконуючи зміну кількості класів можливо визначити групу людей, яка не відноситься до абсолютно здорових і вважати, що вона знаходиться на якійсь відстані від них, при цьому не відноситься до хворих. Проблемою даного підходу є те, що він потребує конкретних ознак, притаманних кожній групі, а значить необхідно задавати норму, якої дотримуються вже багато років, а це суперечить завданню роботи.

Кластеризація – розподіл по групах таким чином, щоб представники однієї групи були схожі між собою більше, ніж схожі на представників іншої. Задача кластеризації виконується автоматично, не залежно від наявності конкретних ознак «схожості». Ознаки можуть обиратись алгоритмом самостійно, а їх внесок в остаточний результат буде змінюватись з кількістю вхідних даних. Завдання кластеризації відноситься до класу задач без вчителя. Тобто ми не знаємо наперед результат, який отримаємо. Алгоритм сам обирає вагу кожного вхідного значення спираючись на мінімальну похибку результату. Завдання людини полягає в тому,

щоб налаштувати параметри програми таким чином, щоб практично ідентичні значення потрапляли до одного кластеру. Методи кластеризації вирішують завдання роботи наступним чином: проводяться налаштування так, щоб виникли деякі великі кластери (норма) та малі (відхилення від норми). Оцінити величину відхилення від норми можна дослідивши на скільки різні між собою «нормальні» кластери та «викиди». Для цього доволі зручно використовувати ієрархічну (агломераційну) кластеризацію [23], реалізовану в бібліотеці мови програмування Python – scikit-learn. Результатом є дерево, яке визначає окремі кластери, потім об'єднує їх в більші і т.д (рисунок 2.4).

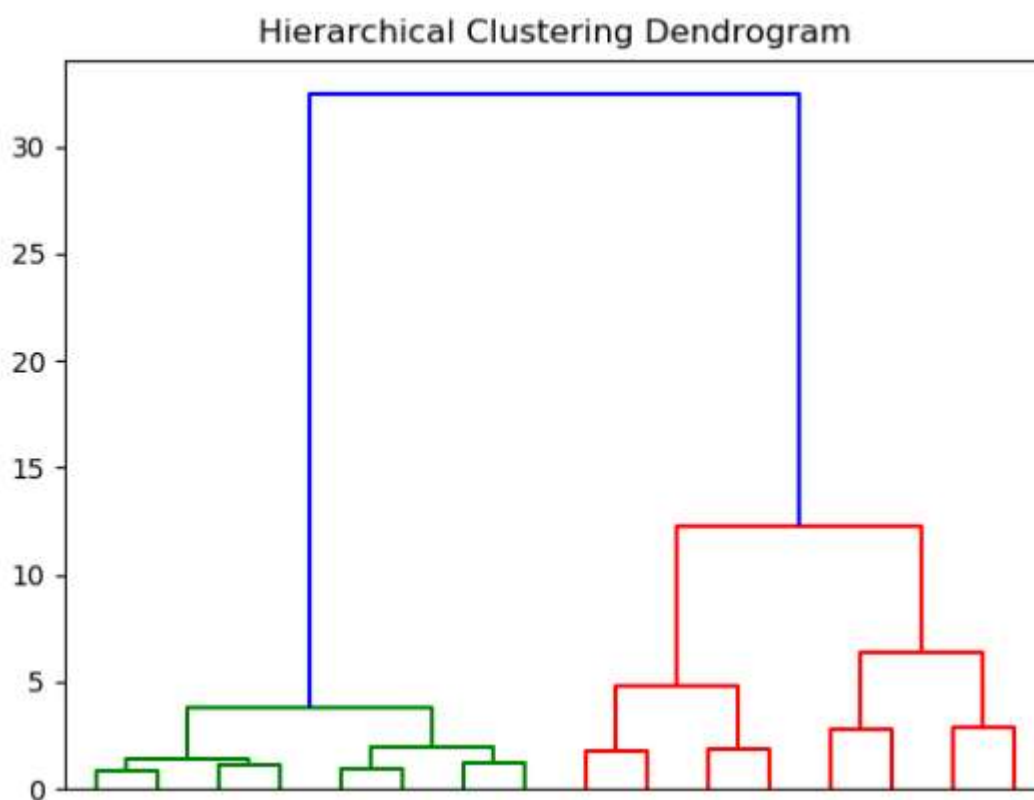


Рисунок 2.4

На рисунку 2.4 можна умовно визначити дерево з двома великими «гілками» (зелена та червона) , які в свою чергу розділяються на менші і закінчуються «листя», що являються найменшими групами вхідних даних (в даній роботі це можуть бути окремі пацієнти). Враховуючи, що більшість пацієнтів, чий сигнали будуть використані, здорові, можливо зробити висновок, що найменші групи з

рівня «листя» – аномалії, а їх відстань від норми оцінюється кількістю та положенням ієрархій, що відділяють їх від основного дерева.

Зазначимо, що класифікація та кластеризація в поставленій задачі тим чи іншим чином зводяться до пошуку аномалій, тому використання методів, які розроблені саме для розв’язку таких задач цілком виправдано.

Пошук аномалій – задача пошуку об’єкту, який відрізняється від об’єктів вибірки. Даний тип задач розділяється на два напрямки : детектування викидів (Outlier Detection) і «новизни» (Novelty Detection). Як і викид, новий об’єкт – об’єкт, який відрізняється, але на відміну від викиду він відсутній в тренувальній вибірці (він з’явиться через деякий час, і задача стоїть в тому, щоб визначити його при появі). Існують деякі особливості, які відрізняють методи кластеризації та класифікації від пошуку аномалій, наприклад, дисбаланс класів.

Викиди являються наслідком:

1. Помилки в даних (неточні виміри, округлення, невірні записи, тощо);
2. Наявності шумових об’єктів (невірно класифікованих);
3. Наявності об’єктів «інших» вибірок (наприклад, показники датчика, що вийшов з ладу).

Існує декілька основних методів виділення викидів:

1. Статистичні тести;
2. Модельні тести;
3. Ітераційні методи;
4. Метричні методи;
5. Методи підміни задачі (використання класифікації чи кластеризації);
6. Методи машинного навчання;
7. Поєднання методів в ансамблі.

Статистичні тести, як правило, використовують для визначення екстремальних значень (рисунок 2.5). Для цього використовують, наприклад, Z-value (2.1) або Kurtosis measure (2.2) методи:

$$Z_i = \frac{|x_i - \mu|}{\sigma}, \quad (2.1)$$

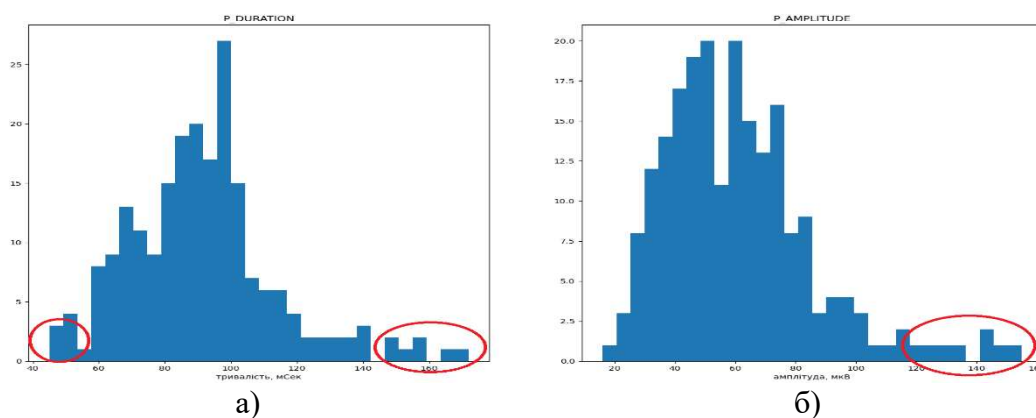
x_i – i -те значення;

μ – середнє значення;

σ – стандартне відхилення.

$$Z_i = \frac{1}{n} \sum_{i=1}^n x_i^4, \quad (2.2)$$

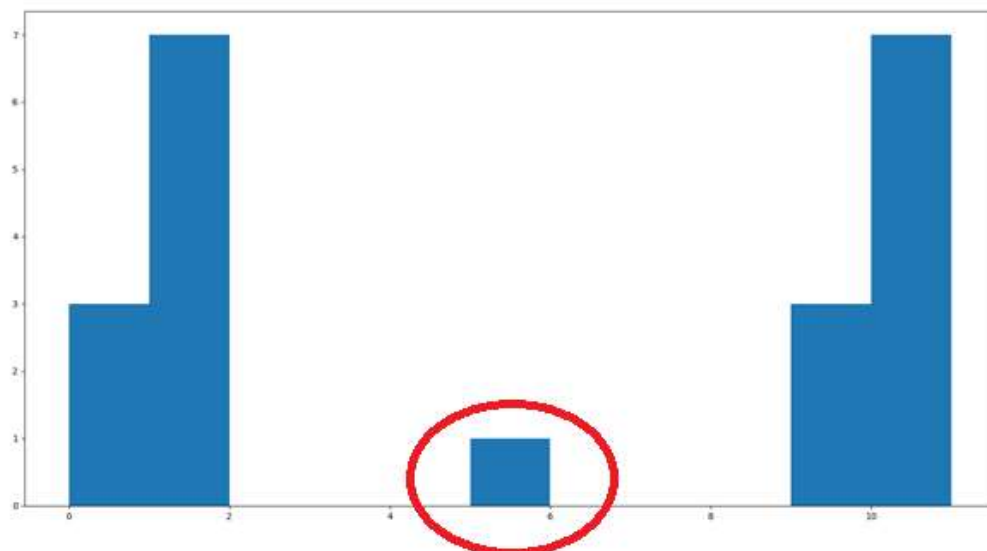
n – кількість значень x_i .



Червоними областями виділені екстремальні статистичні викиди для даної вибірки.

Рисунок 2.5 – Розподіл значень тривалості зубця Р (а) та його амплітуди (б) для першого відведення серед усіх пацієнтів з використаних в роботі даних

Недоліком статистичних методів є виділення лише крайових значень. Це означає, що у випадку, зображеному на рисунку 2.6 методи не можуть вказати на викид.



Червоною областю вказаний викид відносно двох вибірок.

Рисунок 2.6

Ідея модельних тестів полягає у побудові моделі, яка описує дані. Точки, що сильно відхиляються від моделі (на яких модель сильно помиляється) і є аномаліями. При виборі моделі слід враховувати природу задачі, функціонал якості і т.д. Такі методи добре працюють для пошуку «новизни», але не підходять для пошуку викидів, адже спочатку необхідно зробити модель, для якої необхідна вибірка без викидів, тобто необхідно прибрати те, що в контексті поставленої задачі необхідно знайти.

Ітераційні методи на кожній ітерації «відшаровують» групу «особливо підозрілих». Наприклад, в n -мірному просторі можливо видаляти випуклу оболонку точок-об'єктів, приймаючи їх за викиди. Дані методи мають один великий недолік – час виконання алгоритмів великий. Враховуючи, що на кожній ітерації алгоритм має аналізувати оболонку n -розмірної поверхні, програма може виконувати обрахунку протягом дуже великого часу.

Метричні методи одні з найпопулярніших. В них прописується існування деякої метрики в просторі об'єктів, яка допомагає знайти аномалії. Інтуїтивно зрозуміло, що у викиду мало сусідніх точок в деякій відстані, а у типовій точки багато. Існує велика кількість метрик, наприклад відстань Махаланобіса [24].

В даній метриці використовується припущення, що набір точок описує деякий простір. Знаходиться центр маси даного простору та обчислюється відстань кожної точки до знайденого центра. Визначення викидів схоже на статистичний метод *Z-value*. Основним недоліком є те, що в нього закладене подібне припущення як і в *Z-value*: точки описують сферичну поверхню. Тобто, якщо точки матимуть еліптичний розподіл, то частина може вважатись викидами. Ще більш складна ситуація в тороподібних поверхнях, коли центр маси знаходиться поза поверхнею.

Задача пошуку аномалій вже давно сприймається окремою від таких розділів машинного навчання, як класифікація та кластеризація. Для прикладу візьмемо метод опорних векторів (*OneClassSVM*), *IsolationForest*, еліпсоїдальна апроксимація даних (*EllipticEnvelope*). Реалізація даних методів представлена в *scikit-learn*.

Метод опорних векторів – набір алгоритмів машинного навчання з вчителем. Основна ідея – перехід в простір вищої розмірності та пошук гіперплощини з максимальним зазором в цьому просторі. По обидвом сторонам будується ще 2 паралельні гіперплощини. Алгоритм припускає, що найбільша різниця або відстань між гіперплощинами забезпечують мінімальну середню помилку класифікатора [25]. Основний недолік методу для розв’язання поставленої задачі – необхідність наявності вчителя, метод краще використовувати в задачах *novelty detection*.

Еліпсоїдна апроксимація даних формує «хмару» точок, які моделюються у вигляді внутрішньої площі еліпсу. Метод виключно добре працює для нормально розділених даних на площині, ступінь викиду, фактично, визначається як в *Z-value* з усіма характерними йому недоліками. До того ж, використання n – мірного простору для цього методу є не коректним.

Isolation Forest – один з найпопулярніших методів для вирішення задач пошуку аномалій, є ансамблем дерев [26]. Суть полягає в тому, що при випадковій побудові дерев викиди будуть потрапляти в «листя» на ранніх етапах (на не великій глибині дерева), тобто їх простіше ізолювати. Дерево будується доки кожен об’єкт

не буде належати окремому «листу».

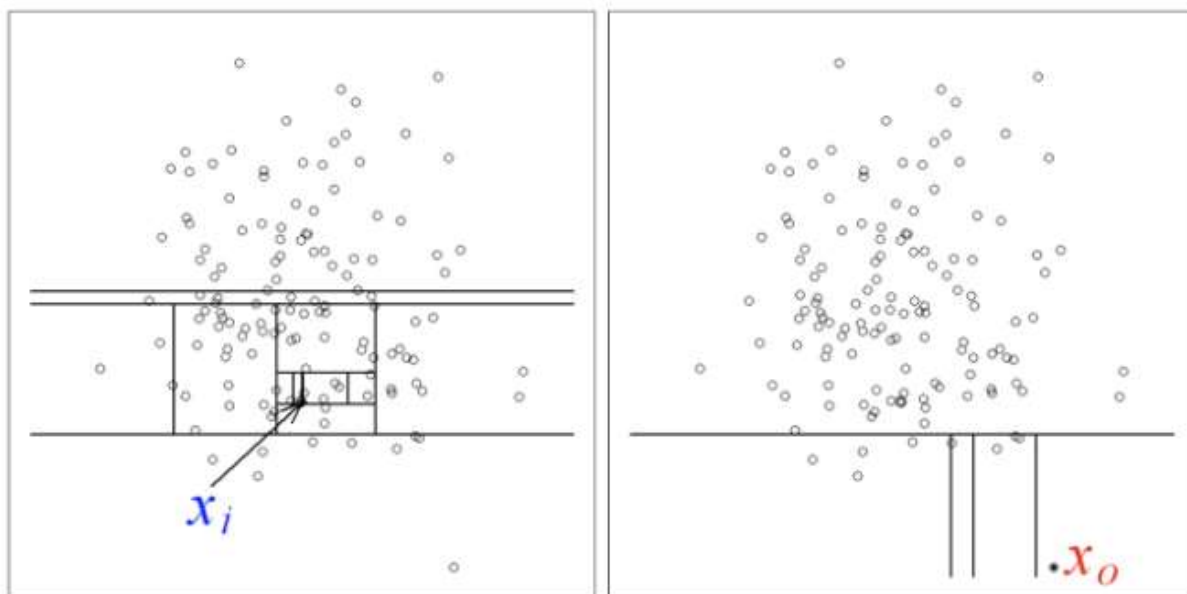


Рисунок 2.7 – Визначення нормальних значень та викидів

Як видно з рисунку 2.7 викид (X_0) потребує лише кількох побудов для присвоєння йому «листа», в той час як нормальне значення знаходиться в складній системі «гілок». Ступінь нормальності знаходиться як:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}, \quad (2.3)$$

де $h(x)$ – довжина шляху спостереження;

$c(n)$ – середня довжина невдалого шляху в дереві двійкового пошуку;

n – кількість зовнішніх вузлів.

Для багатовимірних задач шлях проводиться до координати вектору, що описує один об'єкт. Isolation Forest адекватно працює з багатовимірними задачами. Автор методу зазначає, що він точніший та швидший за методи локального рівня викиду та Ordinal Regression and Classification Algorithms (ORCA, набір методів для сериовища MATLAB), що важливо для великих масивів даних, що збільшують кінцеву точність. Також не потребує задання метрики або іншої апріорної

інформації про дані та здатен розпізнати аномалії різних видів : як ізольовані точки з низькою локальною густиною так і кластери аномально малих розмірів.

Універсальність даного методу – одна з основних переваг, водночас і з недоліків. Тобто, знаючи специфіку даних, ідеальну для них метрику та маючи тренувальні вибірки можливо підібрати метод, який буде точніший за Isolation Forest.

Порівнявши доступні та розповсюджені методи, враховуючи їх переваги і недоліки, зважаючи на специфіку даних було обрано метод машинного навчання Isolation Forest, реалізований в scikit-learn [27].

Існує ряд функцій, які необхідно використовувати для побудови програми:

1. `decision_function(self, X)` – повертає середні значення відхилення кожного об'єкту від базових кластерів;
2. `score_samples(self, X)` – оцінка аномальності кожного об'єкту;
3. `fit(self, X[, y, sample_weight])` – створює або змінює сітку вагових коефіцієнтів (тренує нейронну мережу);
4. `predict(self, X)` – проводить об'єкти через мережу, визначає їх відстань від базисних кластерів, на основі яких визначає належність об'єкту до нормальних значень (inlier) або аномалій (outlier);
5. `fit_predict(self, X[, y])` – виконує дві попередні функції на одних і тих самих даних, тобто проводить пошук викидів, а не аномалій;
6. `get_params(self[, deep])` – повертає параметри мережі;
7. `set_params(self, **params)` – змінює параметри мережі.

Параметри методу Isolation Forest:

1. `n_estimators` – кількість базових оцінщиків;
2. `max_samples` – кількість вибірок з вхідних даних для тренування кожного оцінщика;
3. `Contamination` – частина викидів, використовується для точного налаштування меж базису;
4. `max_features` – максимальна кількість використаних ознак;

5. `bootstrap` – доповнення вибірки додатковими випадковими деревами при тренуванні;
6. `n_jobs` – кількість паралельних процесів;
7. `random_state` – керує псевдовипадковістю вибору об'єкту та значень розділення для кожного кроку створення нової гілки;
8. `verbose` – керує багат шаровістю побудованого дерева;
9. `warm_start` – якщо `True`, доповнює вже створене дерево, якщо `False`, то починає підгонку спочатку.

2.4 Висновки до розділу 2

Розглянуті методи машинного навчання, їх використання та можливості. На основі знань про особливості, недоліки і переваги розглянутих методів був обраний алгоритм Isolation Forest, яким надалі буде розв'язуватись задача даної дипломної роботи. Розглянуті його параметри та функції, що будуть використовуватись при постановці експерименту (Розділ 3).

3 ЕКСПЕРИМЕНТАЛЬНА ЧАСТИНА

3.1 Опис даних

Експериментальні дані для роботи – це база даних кардіологічних досліджень курсантів військової академії в Центральному військовому шпиталі. Вона налічує ЕКГ 139 чоловік, віком 17 – 19 років на момент вимірювань. ЕКГ представляє собою сигнали частотою дискретизації $f_s = 1000$ Гц, тривалістю 2,5 хвилин, шести основних відведень : I, II, III, aVR, aVL, aVF (див. розділ 1). Також наявні додаткові відведення PC1, PC2, PC3, [PC], які математично розраховані з основних.

База даних може бути вчитана з допомогою Cardiol. Дана програма дозволяє відкрити базу, продивитись відведення, отримати усереднену ЕКГ (рисунок 3.1) та її параметри для кожного основного відведення (таблиця 3.1).

Таблиця 3.1 – Параметри усередненої ЕКГ

№	Позначення параметру	Значення параметру
1	PR_DURATION	тривалість інтервалу PR
2	P_DURATION	тривалість зубця Р
3	P_AMPLITUDE	амплітуда зубця Р
4	P_NOTCHED	наявність подвійного зубця Р
5	P1_DURATION	тривалість другого зубця Р за умови його наявності
6	P1_AMPLITUDE	амплітуда другого зубця Р за умови його наявності
7	P_AREA	площа під зубцем Р

Продовження таблиці 3.1

№	Позначення параметру	Значення параметру
8	QRS_DURATION	тривалість QRS комплексу
9	Q_DURATION	тривалість зубця Q
10	Q_AMPLITUDE	амплітуда зубця Q
11	R_DURATION	тривалість зубця R
12	R_AMPLITUDE	амплітуда зубця R
13	S_DURATION	тривалість зубця S
14	S_AMPLITUDE	амплітуда зубця S
15	R1_DURATION	тривалість другого зубця R за умови його наявності
16	R1_AMPLITUDE	амплітуда другого зубця R за умови його наявності
17	S1_DURATION	тривалість другого зубця S за умови його наявності
18	S1_AMPLITUDE	амплітуда другого зубця S за умови його наявності
19	T_AMPLITUDE	амплітуда зубця T
20	T1_AMPLITUDE	амплітуда другого зубця T за умови його наявності
21	J_AMPLITUDE	амплітуда J хвилі
22	ST_MID_AMPLITUDE	середнє значення амплітуди на сегменті ST
23	ST_END_AMPLITUDE	амплітуда в кінці сегменту ST
24	QRS_AREA	площа QRS комплексу
25	TBEG_T_AREA	площа частини зубця T від початку до вершини
26	T_TEND_AREA	площа частини зубця T від вершини до кінця

Продовження таблиці 3.1

№	Позначення параметру	Значення параметру
27	TBEG_T_MAX_DER	найбільше значення похідної на висхідній частині зубця
28	T_TEND_MAX_DER	найбільше значення похідної на низхідній частині зубця
29	QR_DURATION	інтервал Q-R
30	TR_DURATION	інтервал T-R
31	T1R_DURATION	інтервал T1-R за умови наявності T1
32	JT_DURATION	сегмент JT
33	JTA_DURATION	тривалість від точки J до вершини T
34	TA_TEND_DURATION	тривалість інтервалу від вершини до кінця зубця T

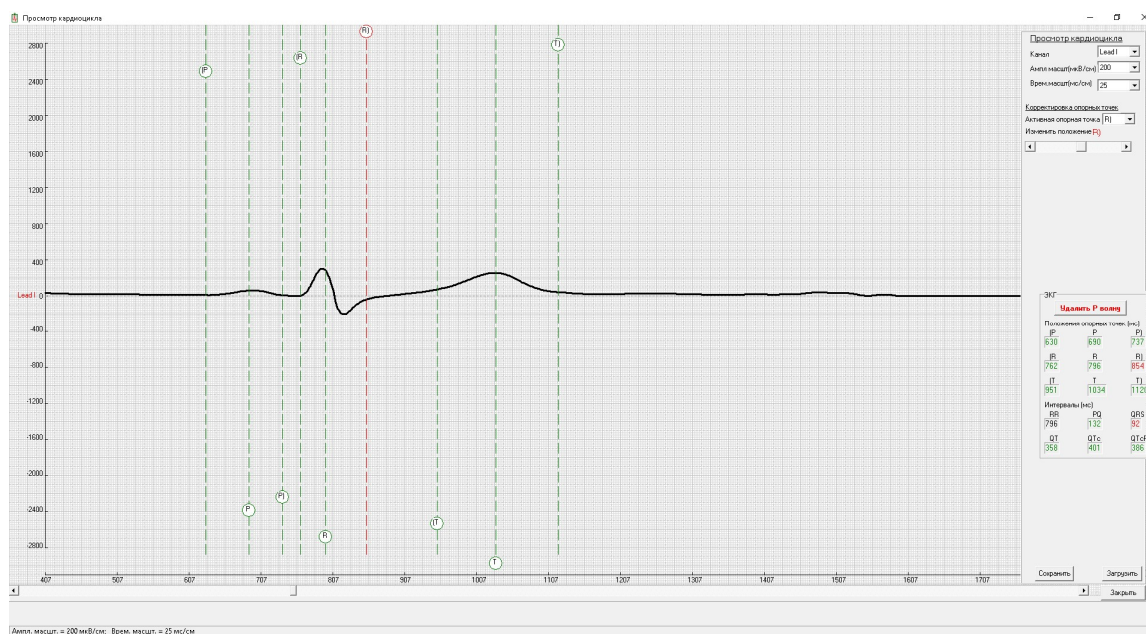


Рисунок 3.1 – Вигляд усередненого кардіокомплексу одного з пацієнтів за першим основним відведенням в програмі Cardiol

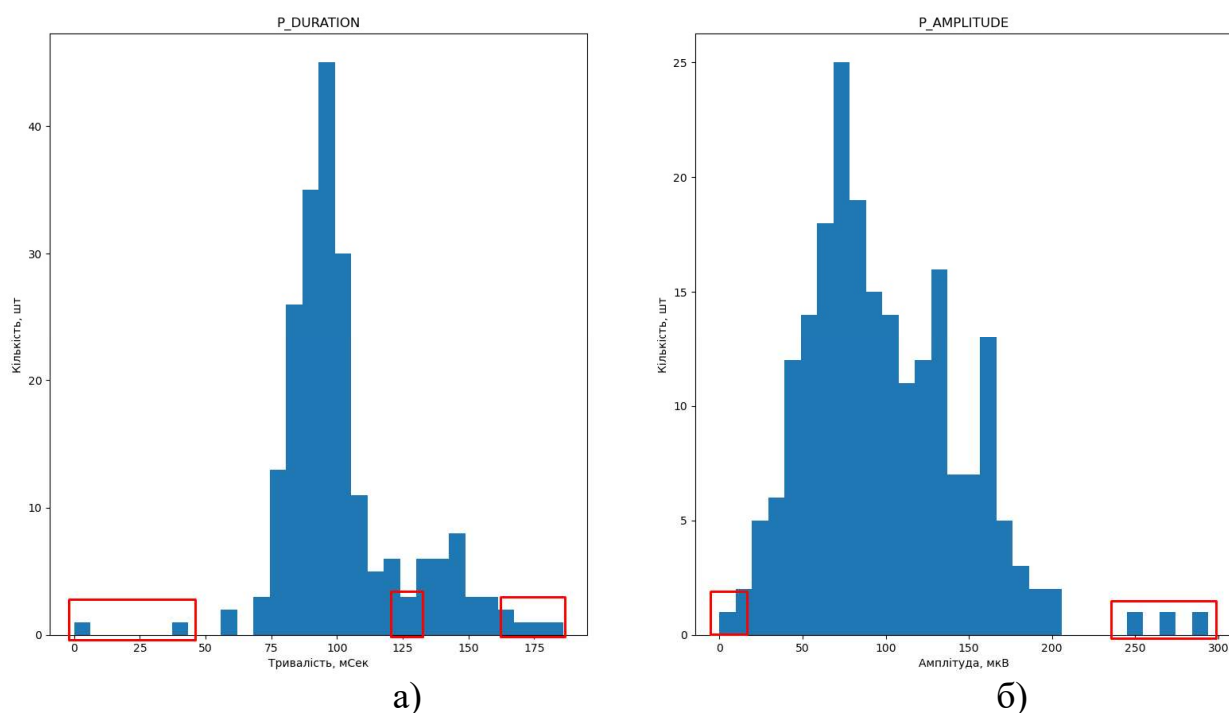
Параметри та усереднений комплекс можна зберегти у окремому файлі формату «.qrs», який може бути відкритий як «.txt». У файлі дані мають кодування XML, тому необхідно створити алгоритм вчитування відліків сигналу ЕКГ та параметрів, записати отримані дані в зручну структуру для подальшого дослідження. Для цього було використані такі бібліотеки мови програмування Python : xml (дозволяє зручно працювати з кодуванням XML), os (необхідно для вибору директорії з даними та отримання назв файлів, які являються унікальним номером кожного виміру) та numpy (використовується для формування масиву чисел). Код створеної функції (opening) представлений в Додатку А. Для зручного формування масиву необхідних параметрів була створена окрема функція (array_formation), код представлений в Додатку А. Обидві функції формують модуль data_preparation.py.

Отримана структура – словник, ключами якого є унікальний номер кожного виміру, значення ключа – ще один словник, ключами якого є назви відведення та параметрів з відповідними їм значеннями.

3.2 Постановка експерименту

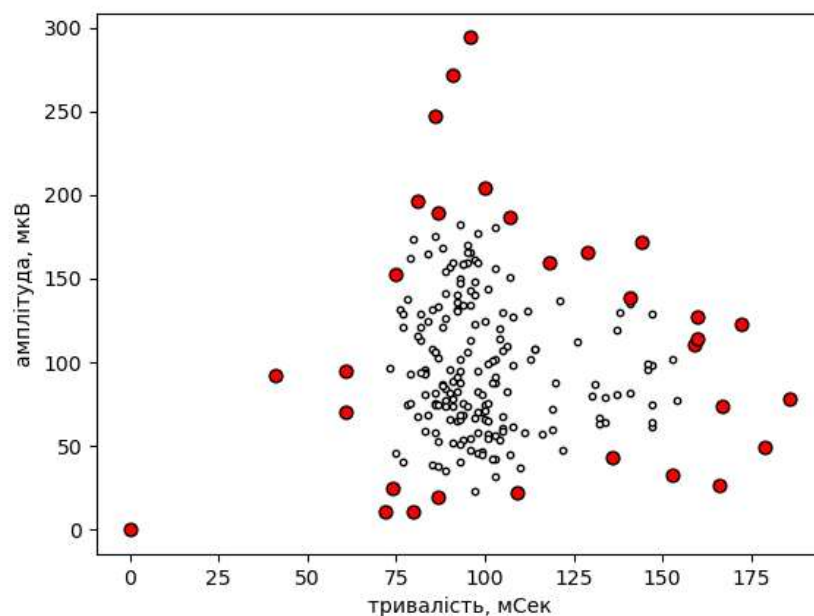
З допомогою створеного модуля (додаток А) необхідно вчитати вхідні дані та використати їх для тренування методу Isolation Forest. Виконаємо ряд дослідів, в кожному досліді відбираємо аномалії. Для цього використаємо обраний метод, представлений у бібліотеці scikit-learn (версія 0.23.1 станом на 1 червня 2020 року)

Дослід 1: 1.1) налаштування методу за замовчуванням (кількість дерев-оцінщиків 100, кількість аномалій обирається методом автоматично, використані всі подані на вхід ознаки), вхідні дані – амплітуда та тривалість зубця Р за другим відведенням (рисунок 3.2).



Червоним зображені зони з наявними викидами

Рисунок 3.2 – Розподіл значень тривалості (а) та амплітуди (б) Р зубця серед робочих даних



Диски білого кольору – точки кожного пацієнта за параметрами тривалості та амплітуди Р. Диски червоного кольору – точки, які були визначені алгоритмом як аномалії.

Рисунок 3.2 – Розподіл даних за параметрами тривалості та амплітуди зубця

Р

1.2) налаштування методу: викиди складають **contamination = 0,1** (10%), кількість дерев $n_estimators = 50$, вхідні дані – амплітуда та тривалість зубця Р за першим відведенням.

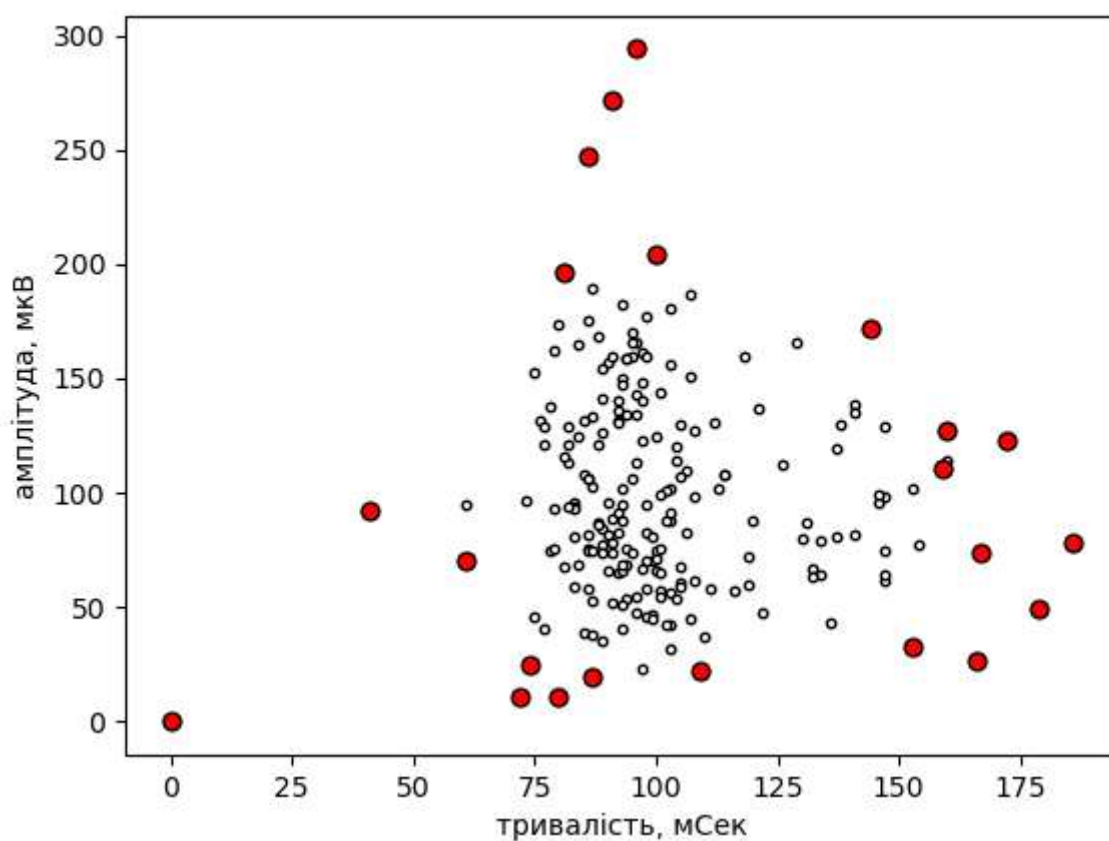


Рисунок 3.3

Даний дослід вказує, що побудована модель працює адекватно, тобто можна візуально оцінити, що алгоритм виділяє точки, що знаходяться далеко від основної вибірки, а значить вирішує поставлене перед ним завдання (рисунок 3.2) та відповідає на зміни параметрів (на рисунку 3.3 наявні 22 аномальні точки, що при округленні складає 10% від кількості вхідних даних). Далі будемо використовувати багатовимірні вектори даних, але відображати будемо за одними і тими ж параметрами, як на рисунку 3.2 для зручності сприйняття.

Дослід 2: 2.1) налаштування методу за замовчуванням, вхідні дані – параметри, що часто використовуються для аналізу ЕКГ за книгою [1] (інтервал Р-Р (рисунок 3.4, а), амплітуда та тривалість зубця Р, тривалість S, сегмент JT,

амплітуда зубця J, параметри сегменту ST(рисунок 3.4, б)). Знайдемо аномалії та їх відстань від побудованої методом норми. Результати наведені в таблиці 3.2 (див. 3.3 Результати).

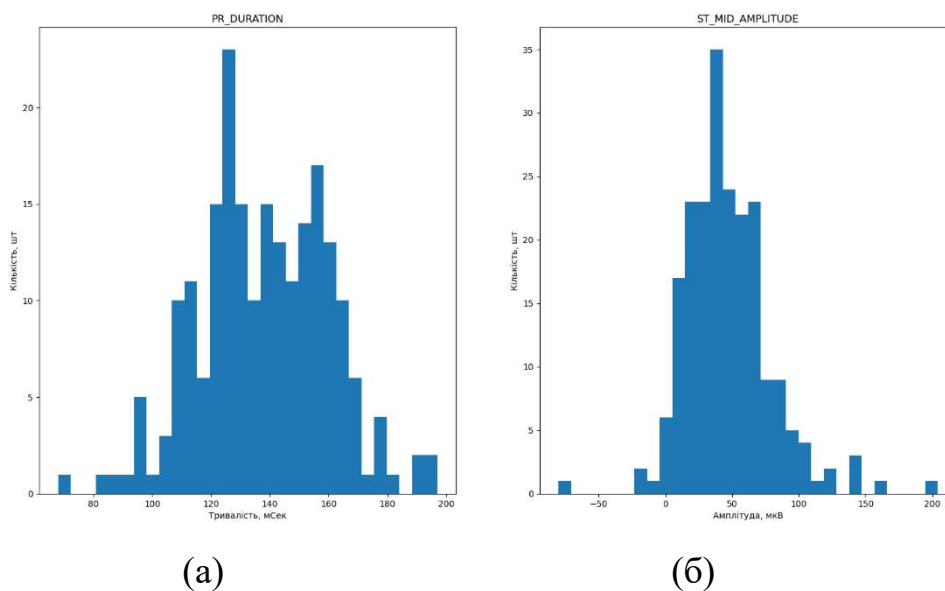


Рисунок 3.4 – розподіл значень P-R інтервалу (а) та середньої амплітуди ST сегменту (а) по вхідним даним

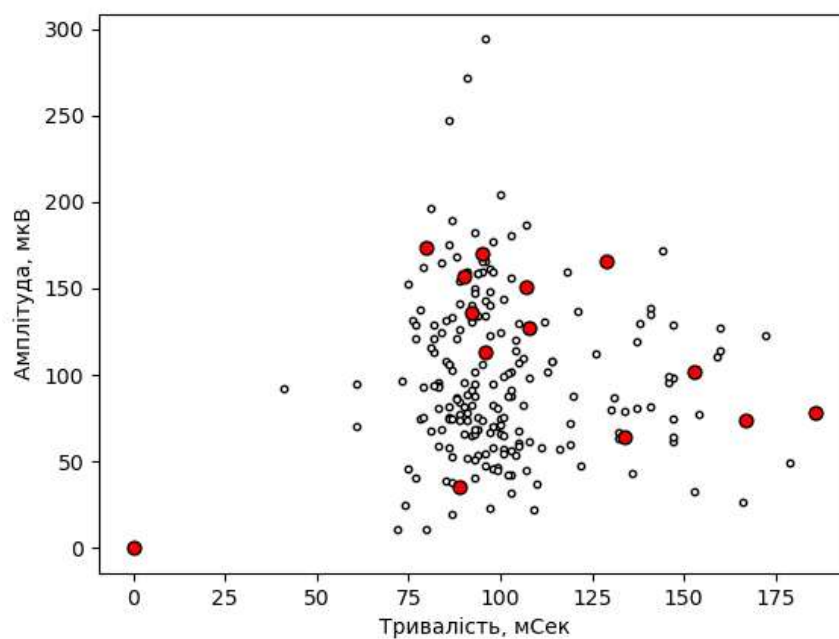


Рисунок 3.5

2.2) налаштування методу: викиди складають $\text{contamination} = 0,1$ (10%), кількість дерев $n_estimators = 50$, інші параметри за замовчуванням. Вхідні дані як в 2.1. Результати наведені в таблиці 3.3 (див. 3.3 Результати).

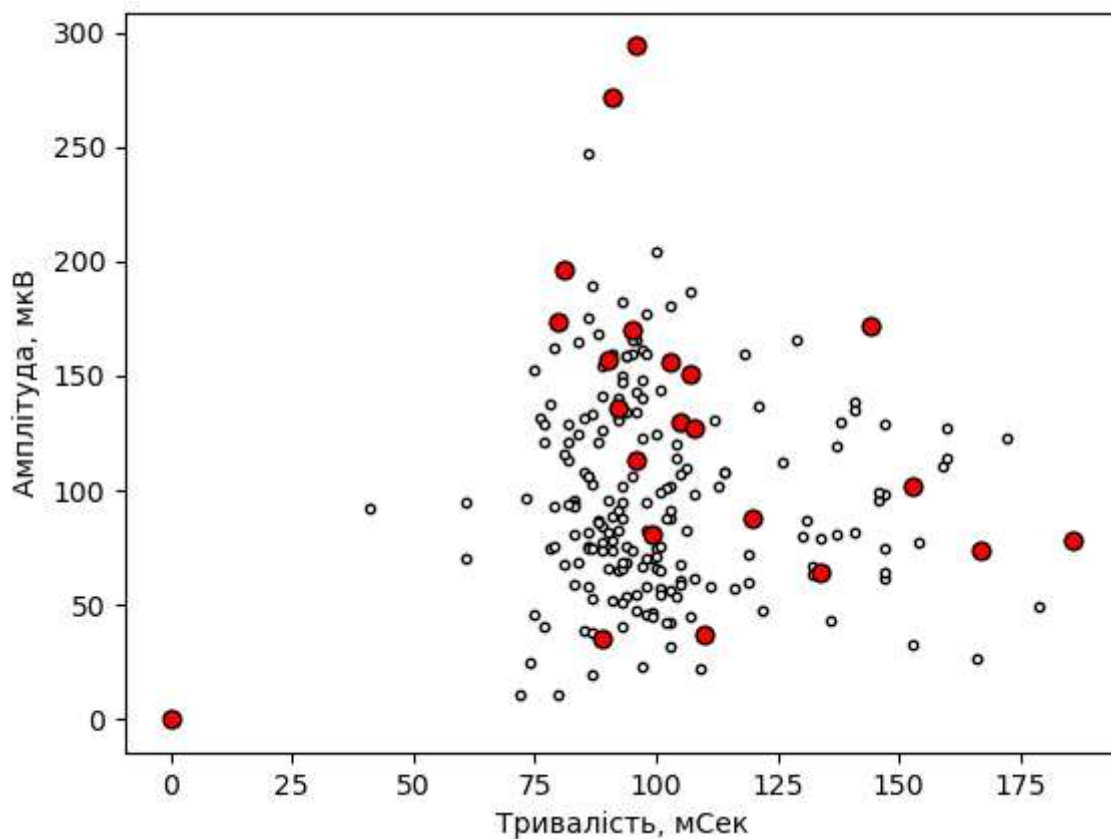


Рисунок 3.6

2.3) налаштування методу: викиди складають $\text{contamination} = 0,05$ (5%), кількість дерев $n_estimators = 200$, інші параметри за замовчуванням. Вхідні дані як в 2.1. Результати наведені в таблиці 3.4 (див. 3.3 Результати).

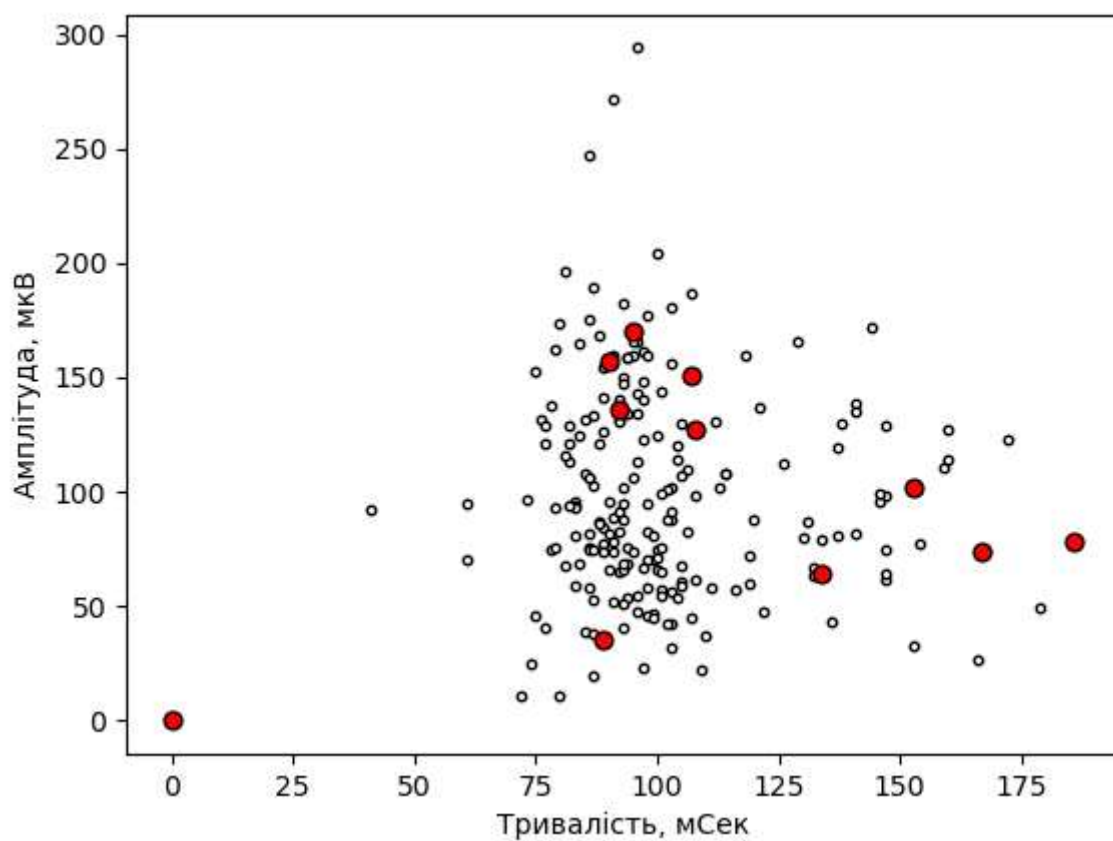


Рисунок 3.6

Дослід 3: 3.1) налаштування методу за замовчуванням, вхідні дані – основні параметри ЕКГ, доступні з використаної бази даних (1 – 4, 7 – 14, 19, 21 – 30, 32 – 34 з таблиці 3.1). Результати наведені в таблиці 3.5 (див. 3.3 Результати).

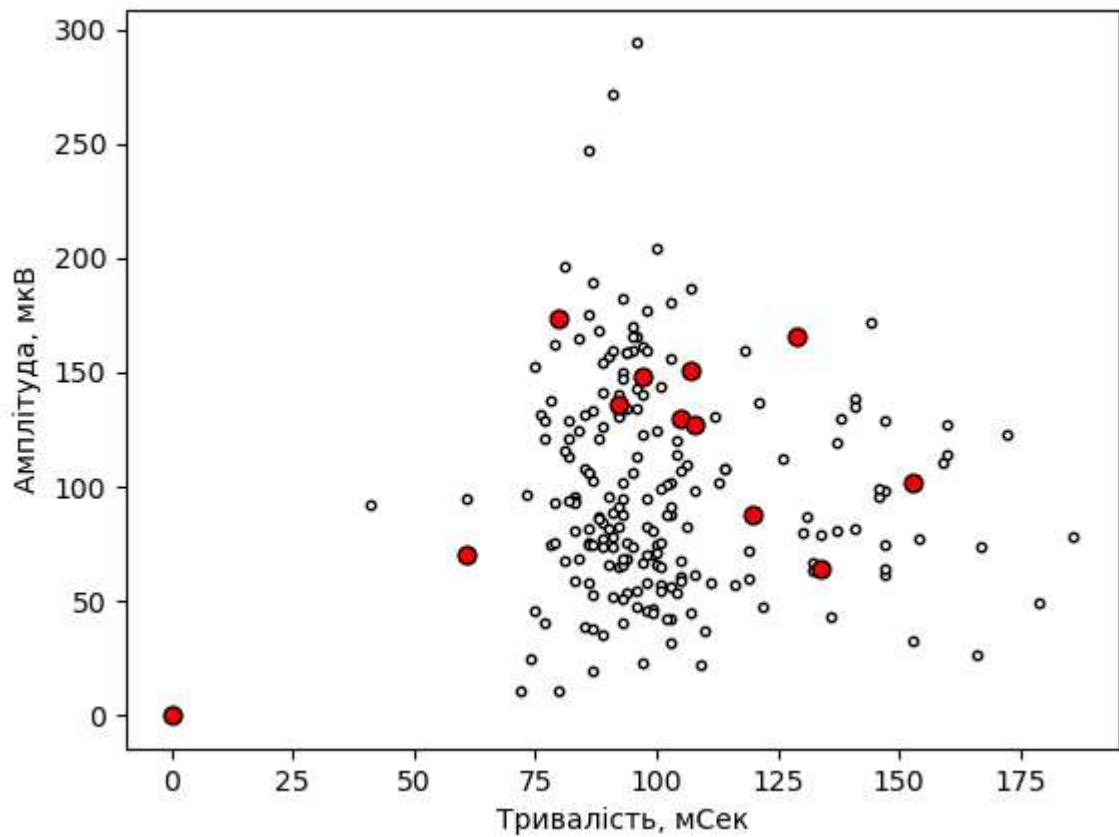


Рисунок 3.7

3.2) налаштування методу: викиди складають $\text{contamination} = 0,1$ (10%), кількість дерев $n_estimators = 50$, інші параметри за замовчуванням. Вхідні дані як в 3.1. Результати наведені в таблиці 3.6 (див. 3.3 Результати).

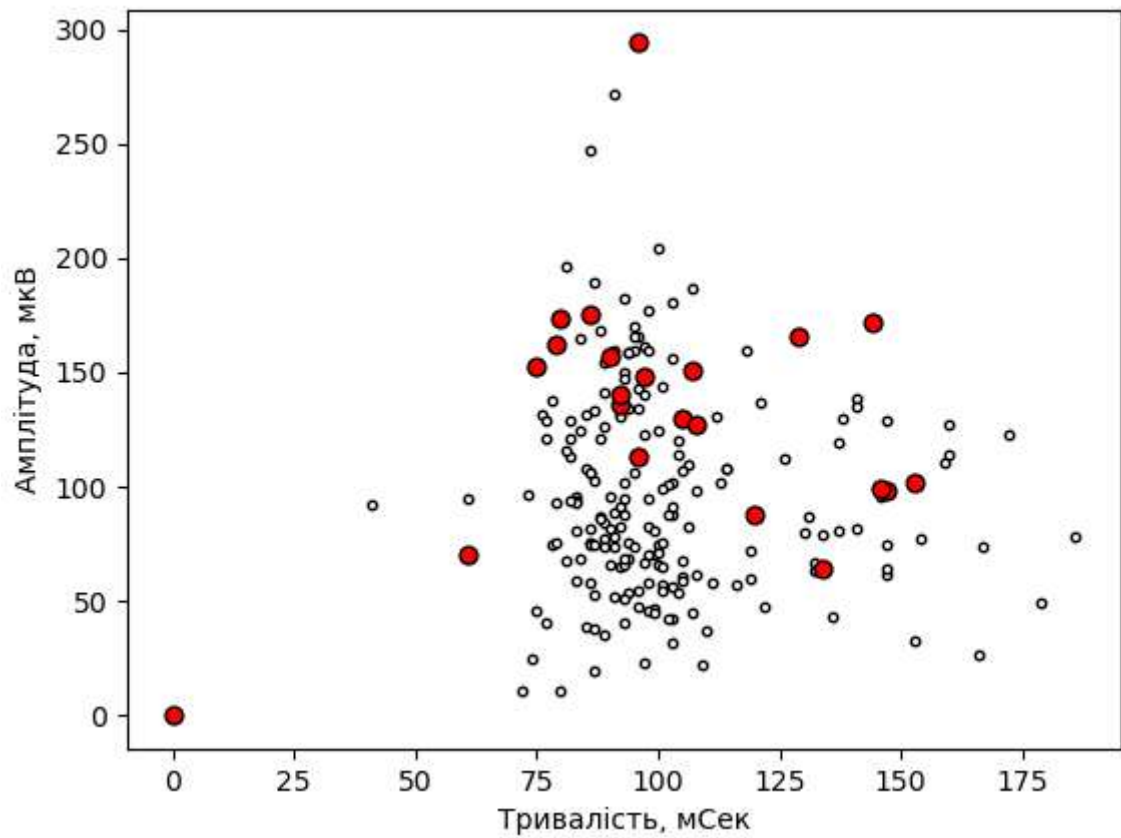


Рисунок 3.8

3.3) налаштування методу: викиди складають $\text{contamination} = 0,05$ (5%), кількість дерев $n_\text{estimators} = 200$, інші параметри за замовчуванням. Вхідні дані як в 3.1. Результати наведені в таблиці 3.7 (див. 3.3 Результати).

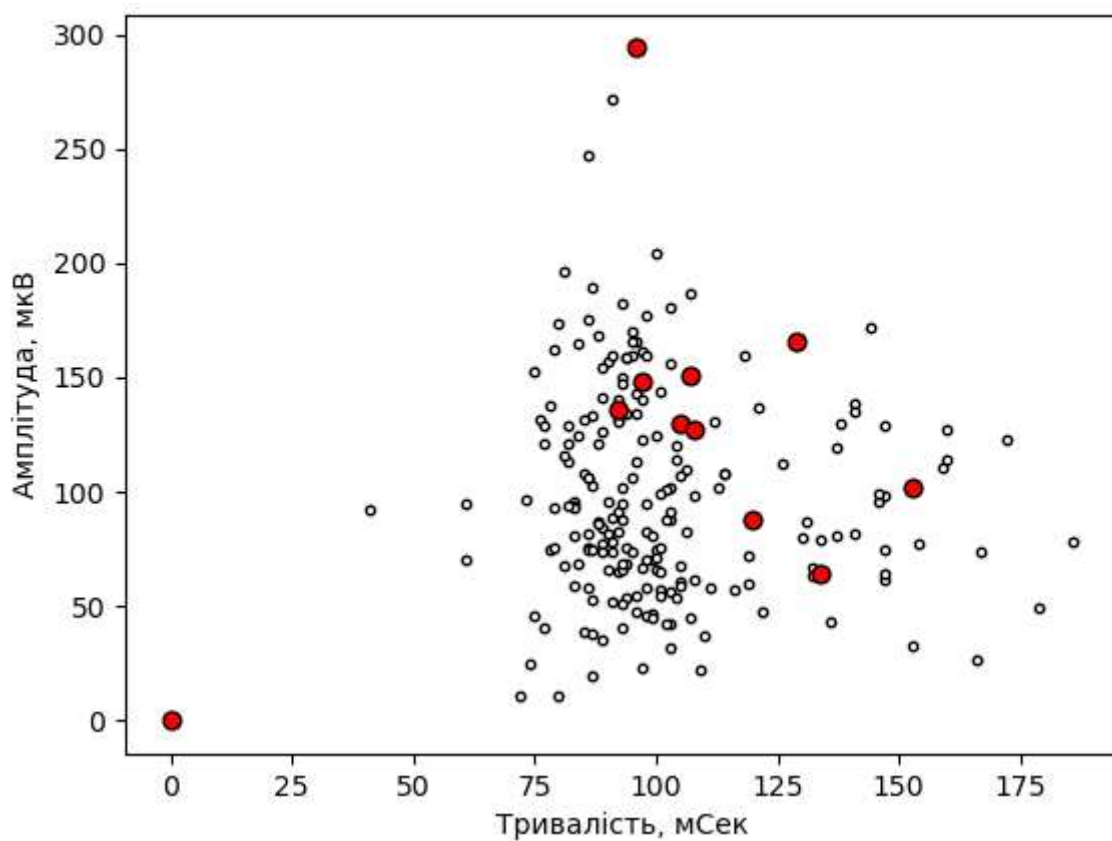


Рисунок 3.9

Дослід 4: 4.1) налаштування методу за замовчуванням, вхідні дані – всі параметри ЕКГ, доступні з використаної бази даних (1–34 з таблиці 3.1). Результати наведені в таблиці 3.8 (див. 3.3 Результати).

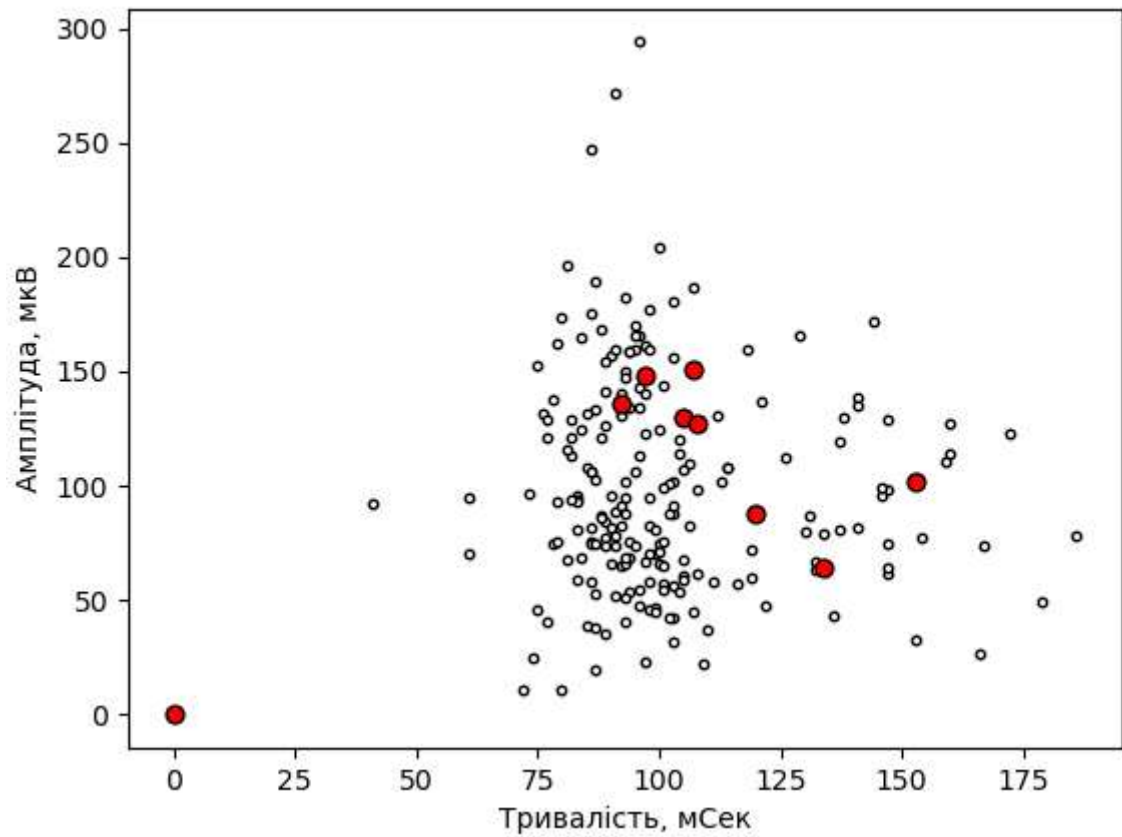


Рисунок 3.10

4.2) налаштування методу: викиди складають $\text{contamination} = 0,05$ (5%), кількість дерев $n_\text{estimators} = 200$, інші параметри за замовчуванням. Вхідні дані як в 4.1. Результати наведені в таблиці 3.9 (див. 3.3 Результати).

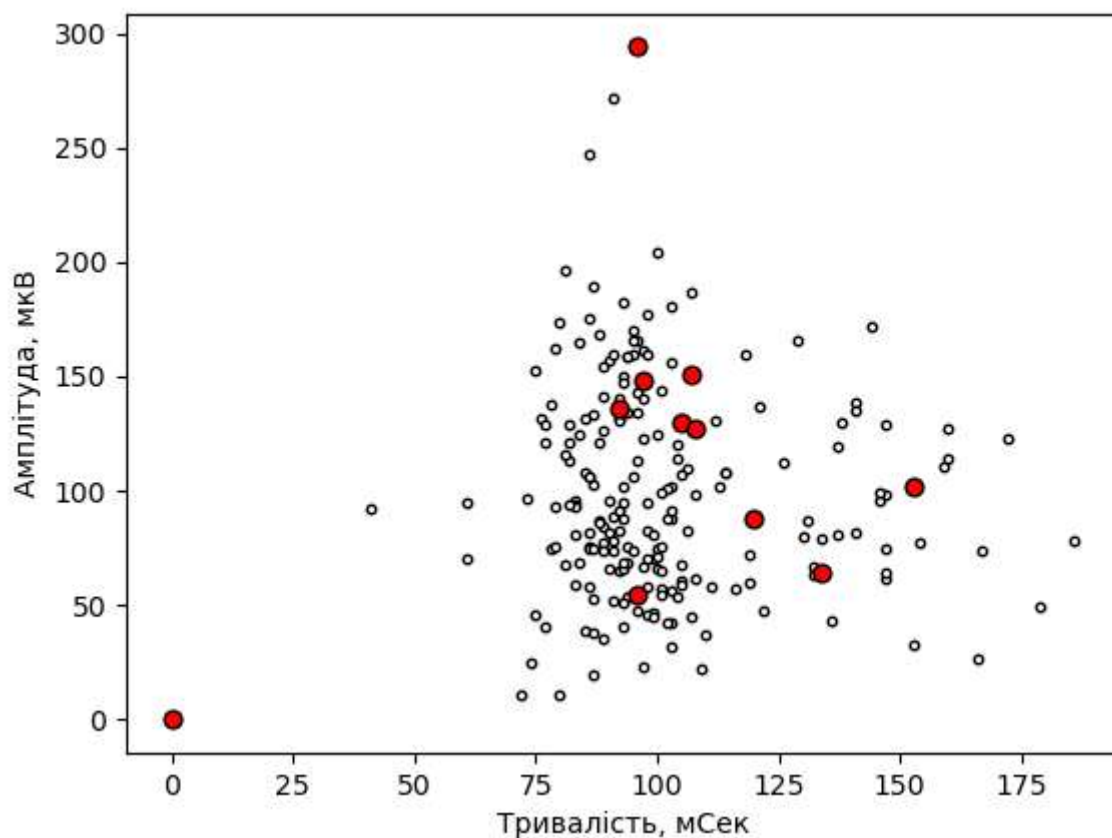


Рисунок 3.11

Дослід 5: налаштування методу: кількість використаних ознак складає $\text{max_features} = 0.8$ (80% від усіх), кількість дерев $\text{n_estimators} = 200$, інші параметри за замовчуванням. Вхідні дані як в 4.1. Результати наведені в таблиці 3.10 (див. 3.3 Результати).

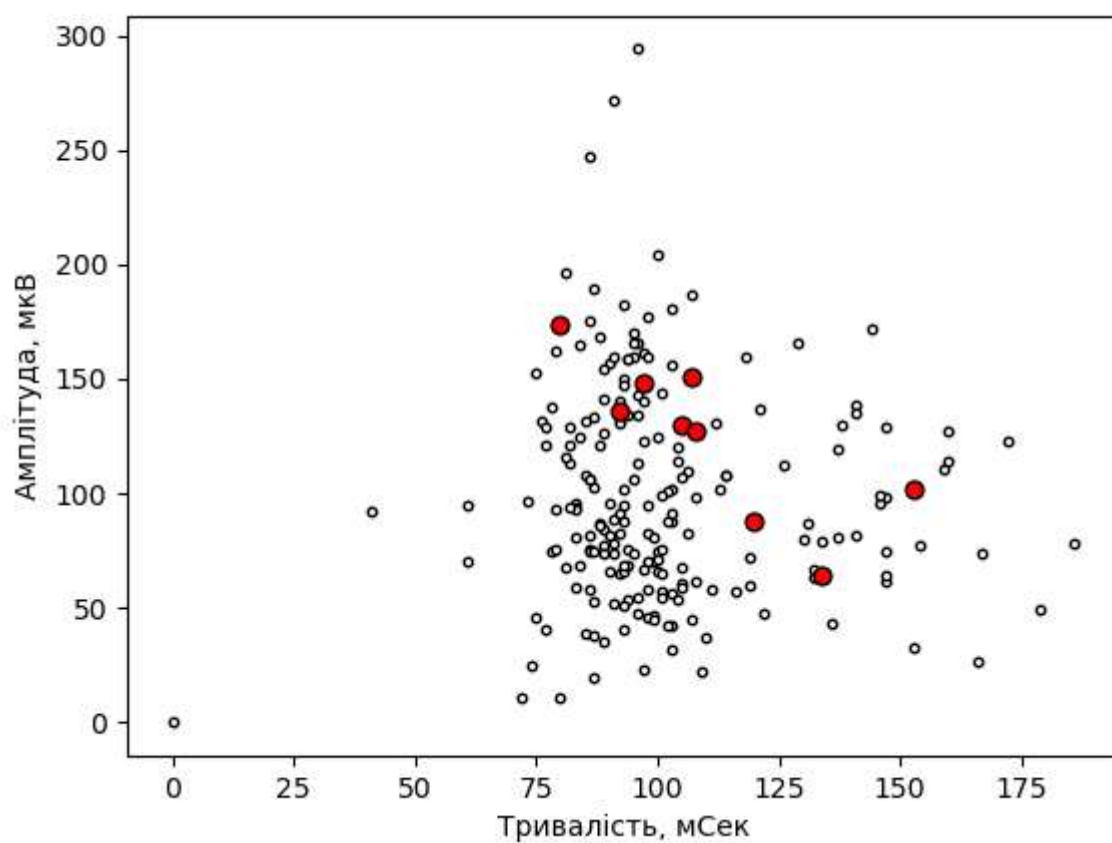


Рисунок 3.12

3.3 Результати

Таблиця 3.2

ID210.qrs	ID104.qrs	ID150.qrs	ID129.qrs	ID121.qrs	ID52.qrs
-0.121251	-0.099227	-0.069495	-0.041721	-0.037284	-0.036170

Продовження таблиці 3.2

ID51.qrs	ID23.qrs	ID124.qrs	ID143.qrs	ID27.qrs	ID65.qrs
-0.034396	-0.015695	-0.014040	-0.012054	-0.008038	-0.005291

Продовження таблиці 3.2

ID207.qrs	ID89.qrs
-0.005243	-0.003182

Таблиця 3.3

ID210.qrs	ID104.qrs	ID150.qrs	ID52.qrs	ID89.qrs	ID221.qrs
-0.203216	-0.122524	-0.112782	-0.092576	-0.073665	-0.057852

Продовження таблиці 3.3

ID51.qrs	ID143.qrs	ID35.qrs	ID197.qrs	ID129.qrs	ID121.qrs
-0.057047	-0.050250	-0.047151	-0.036697	-0.036240	-0.035883

Продовження таблиці 3.3

ID124.qrs	ID207.qrs	ID90.qrs	ID27.qrs	ID151.qrs	ID6.qrs
-0.032928	-0.031412	-0.030938	-0.021280	-0.017920	-0.015948

Продовження таблиці 3.3

ID28.qrs	ID205.qrs	ID108.qrs	ID65.qrs
-0.007051	-0.005382	-0.004621	-0.000537

Таблиця 3.4

ID104.qrs	ID210.qrs	ID150.qrs	ID121.qrs	ID51.qrs	ID52.qrs
-0.123268	-0.091370	-0.051046	-0.045436	-0.043465	-0.035485

Продовження таблиці 3.4

ID207.qrs	ID143.qrs	ID129.qrs	ID124.qrs	ID89.qrs
-0.018295	-0.008600	-0.004641	-0.003634	-0.002987

Таблиця 3.5

ID210.qrs	ID104.qrs	ID51.qrs	ID9.qrs	ID221.qrs	ID151.qrs
-0.098808	-0.097669	-0.055787	-0.052654	-0.029442	-0.026815

Продовження таблиці 3.5

ID119.qrs	ID52.qrs	ID150.qrs	ID23.qrs	ID65.qrs	ID121.qrs
-0.022401	-0.017065	-0.017020	-0.012465	-0.002949	-0.001449

Таблиця 3.6

ID104.qrs	ID119.qrs	ID9.qrs	ID150.qrs	ID210.qrs	ID51.qrs
-0.108664	-0.101980	-0.092775	-0.092403	-0.080672	-0.070738

Продовження таблиці 3.6

ID27.qrs	ID90.qrs	ID221.qrs	ID121.qrs	ID52.qrs	ID23.qrs
-0.066982	-0.057690	-0.055347	-0.055300	-0.049832	-0.031484

Продовження таблиці 3.6

ID108.qrs	ID65.qrs	ID46.qrs	ID129.qrs	ID42.qrs	ID144.qrs
-0.030238	-0.027248	-0.020391	-0.020056	-0.014603	-0.009497

Продовження таблиці 3.6

ID93.qrs	ID198.qrs	ID145.qrs	ID151.qrs
-0.006411	-0.003254	-0.000549	-0.000076

Таблиця 3.7

ID210.qrs	ID104.qrs	ID9.qrs	ID51.qrs	ID150.qrs	ID52.qrs
-0.080690	-0.068158	-0.056808	-0.035925	-0.028675	-0.026957

Продовження таблиці 3.7

ID151.qrs	ID121.qrs	ID221.qrs	ID23.qrs	ID108.qrs
-0.021641	-0.015488	-0.008562	-0.004596	-0.004307

Таблиця 3.8

ID210.qrs	ID9.qrs	ID52.qrs	ID151.qrs	ID150.qrs	ID104.qrs
-0.089616	-0.049772	-0.046034	-0.041115	-0.031042	-0.027241

Продовження таблиці 3.8

ID51.qrs	ID121.qrs	ID221.qrs
-0.026112	-0.019479	-0.016692

Таблиця 3.9

ID210.qrs	ID104.qrs	ID9.qrs	ID150.qrs	ID52.qrs	ID151.qrs
-0.090090	-0.062652	-0.055567	-0.048954	-0.036452	-0.035610

Продовження таблиці 3.9

ID51.qrs	ID108.qrs	ID121.qrs	ID221.qrs	ID4.qrs
-0.013034	-0.011319	-0.009364	-0.006601	-0.002796

Таблиця 3.10

ID210.qrs	ID52.qrs	ID104.qrs	ID9.qrs	ID150.qrs	ID51.qrs	ID151.qrs
-0.053089	-0.050775	-0.046944	-0.043374	-0.042269	-0.030619	-0.007974

Для зручності аналізу отриманих даних таблиці 3.2 – 3.10 відсортовані за величиною віддаленості аномалії від норми.

В досліді 2.1 метод, налаштований за замовчуванням, визначив 14 аномальних векторів значень, що, доволі велике значення для бази даних, яка теоретично має сигнали виключно здорових людей після поверхневого огляду лікарем. Продивившись оцінку сигналів програмою Cardiol стало зрозуміло, що в цілому метод знаходить деякі відхилення від основної маси сигналів, але робить це з деякою помилкою, тому було прийняте рішення зменшити шукану кількість аномалій до 5% та збільшити кількість перевірок та оцінок (дослід 2.3, таблиця 3.4). Наступні досліді проводились з більшою кількістю параметрів. Було припущено, що більша кількість параметрів розширить діапазон норми, а значить, що за її межі має потрапити менше пацієнтів. Метод автоматично визначив 12 аномалій (дослід 3.1), половина з яких були наявні у попередньому досліді. Це означає, що додані параметри значно змінюють норму, яку будує алгоритм. Було зроблено припущення, що необхідно додати усі наявні параметри (дослід 4.1 та 4.2). Алгоритм автоматично визначив 9 аномальних векторів. Це менше 10 відсотків, тому визначаючи більшу кількість, їх відстань від норми змінювалась. Це пов'язано з тим, що частина даних, які раніше вважались нормальними та створювали відповідний діапазон перестали це робити і почали вважатись викидами.

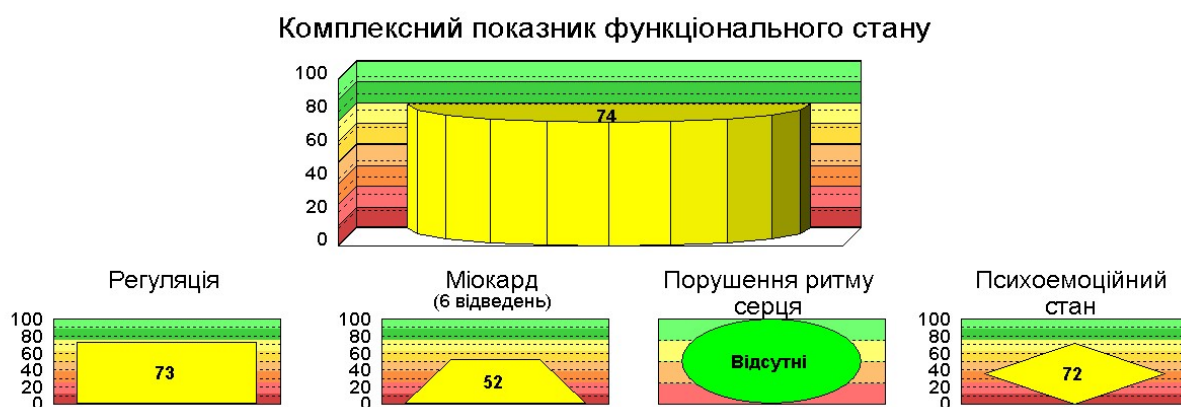


Рисунок 3.13 – Оцінка комплексного показника стану сигналів під номером 210 за допомогою програми Cardiol

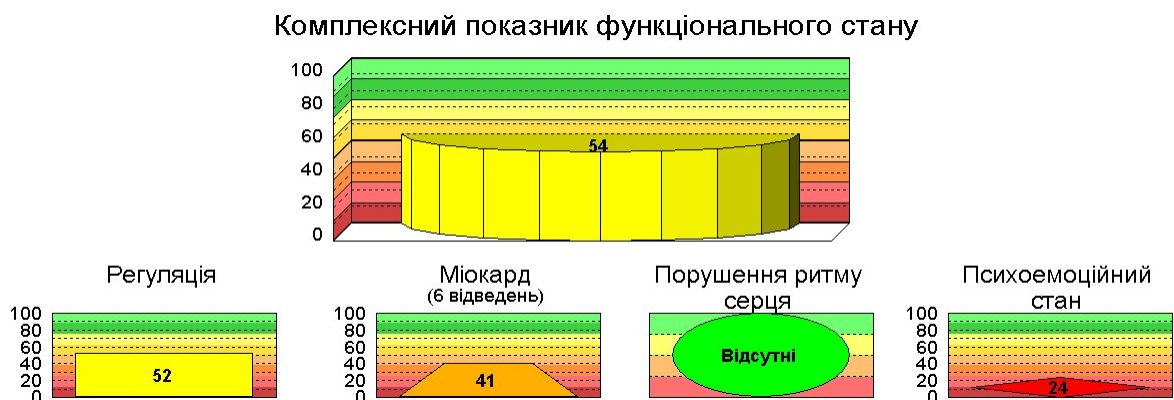


Рисунок 3.14 – Оцінка комплексного показника стану сигналів під номером 104 за допомогою програми Cardiol

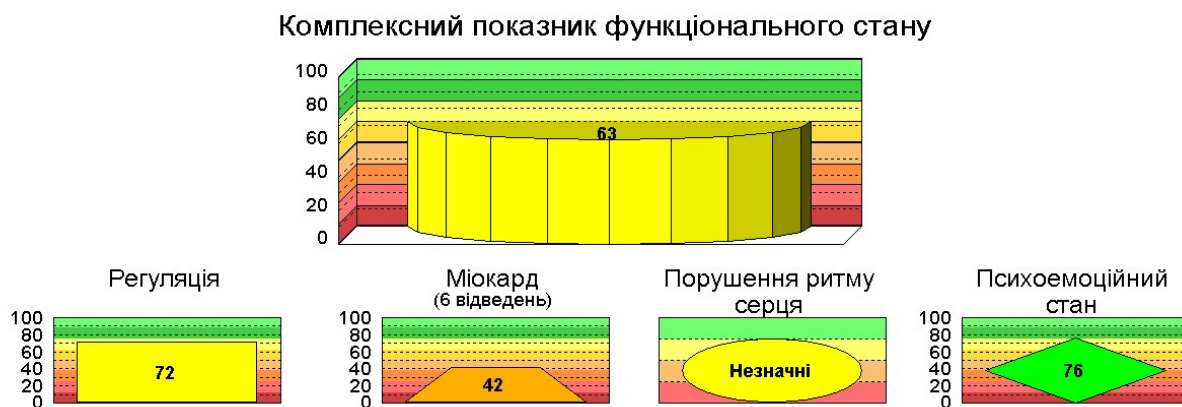


Рисунок 3.15 – Оцінка комплексного показника стану сигналів під номером 9 за допомогою програми Cardiol

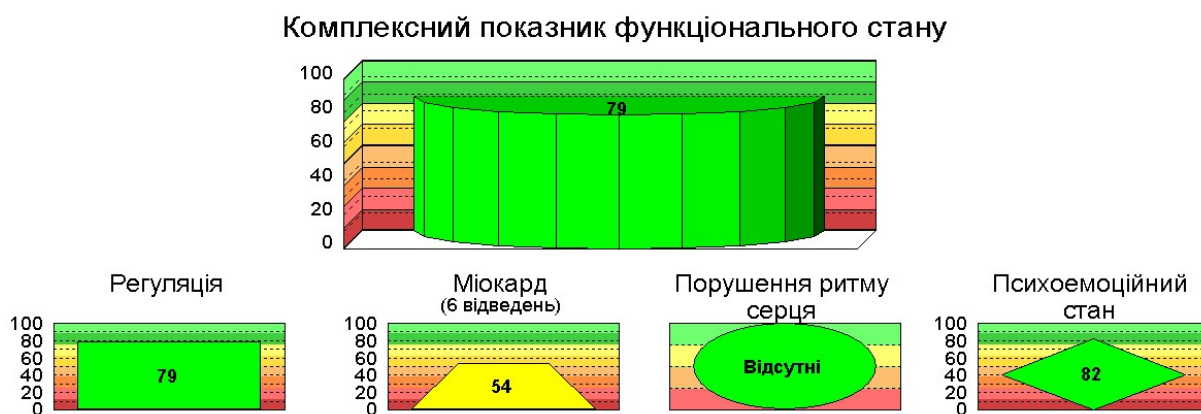


Рисунок 3.16 – Оцінка комплексного показника стану сигналів під номером 150 за допомогою програми Cardiol

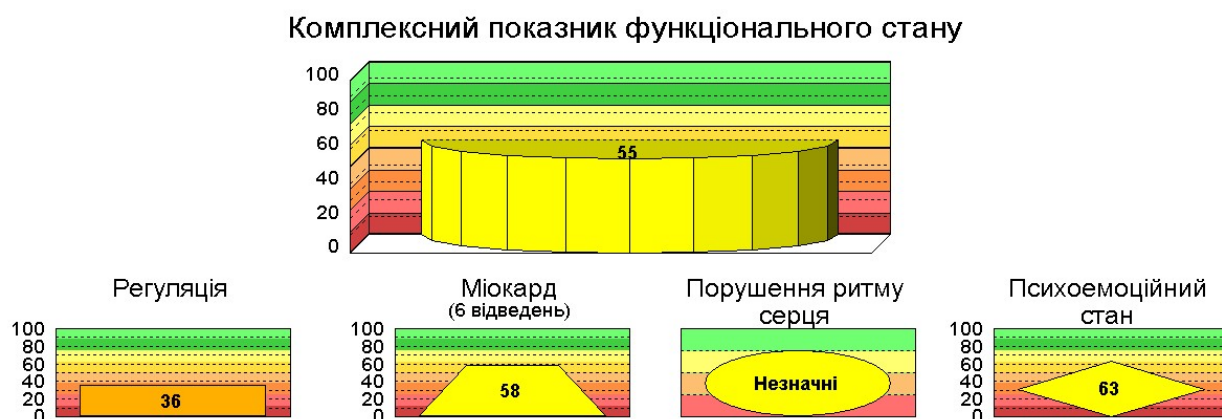


Рисунок 3.17 – Оцінка комплексного показника стану сигналів під номером 52 за допомогою програми Cardiol

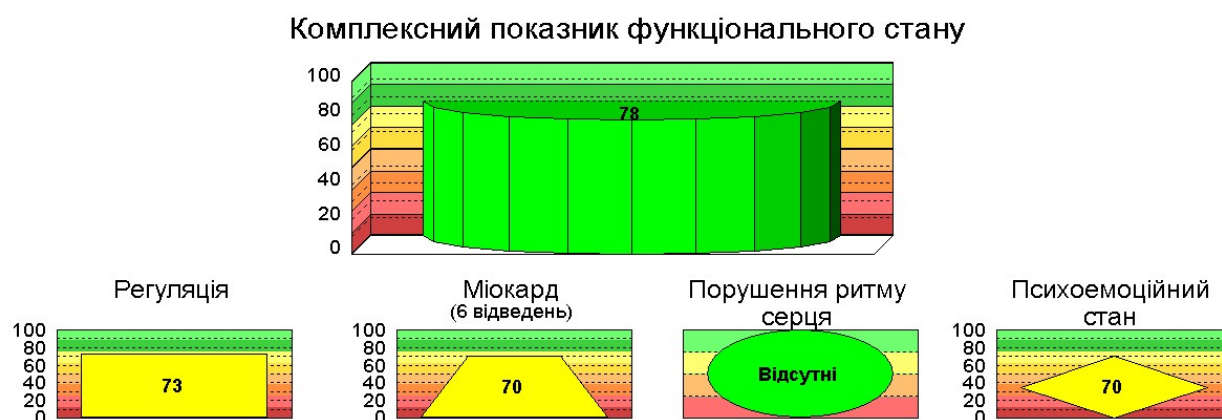


Рисунок 3.18 – Оцінка комплексного показника стану сигналів під номером 151 за допомогою програми Cardiol

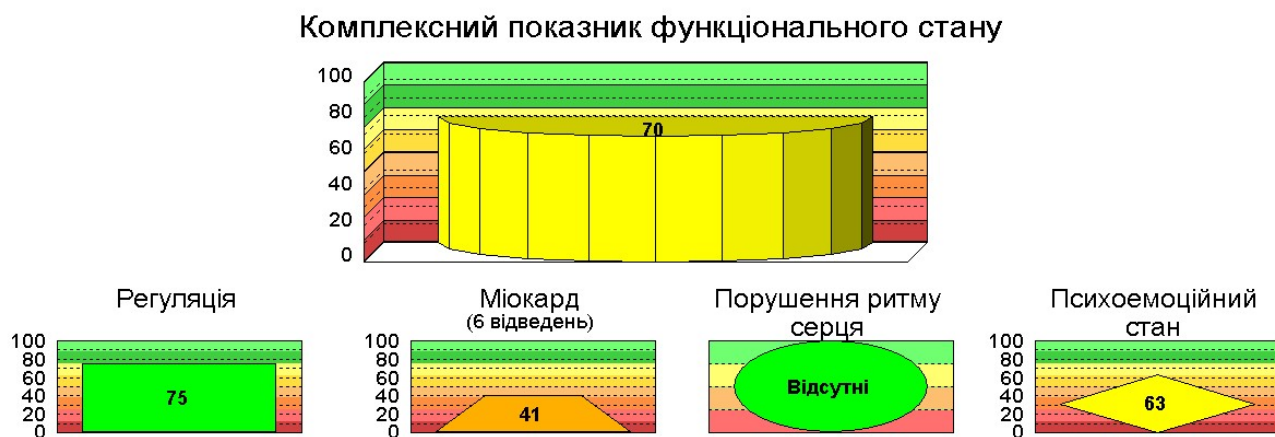


Рисунок 3.19 – Оцінка комплексного показника стану сигналів під номером 51 за допомогою програми Cardiol

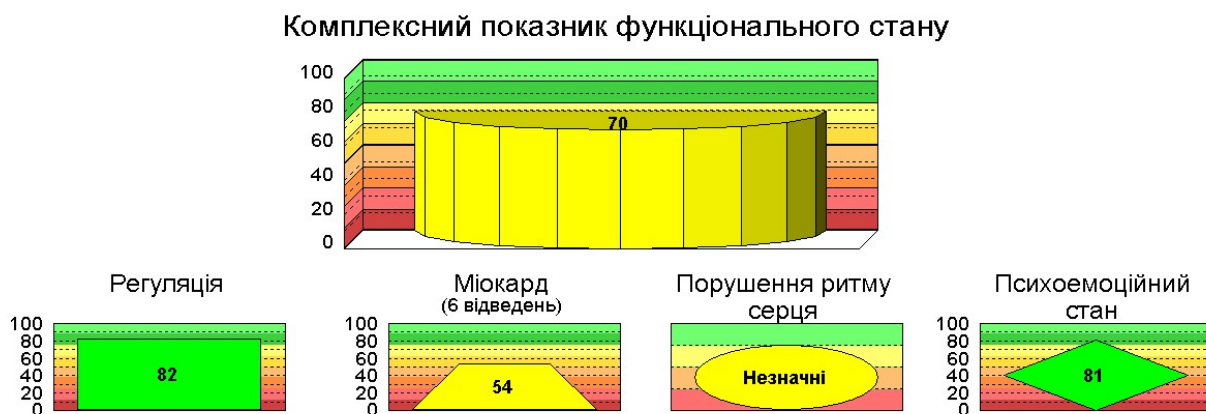


Рисунок 3.20 – Оцінка комплексного показника стану сигналів під номером 108 за допомогою програми Cardiol

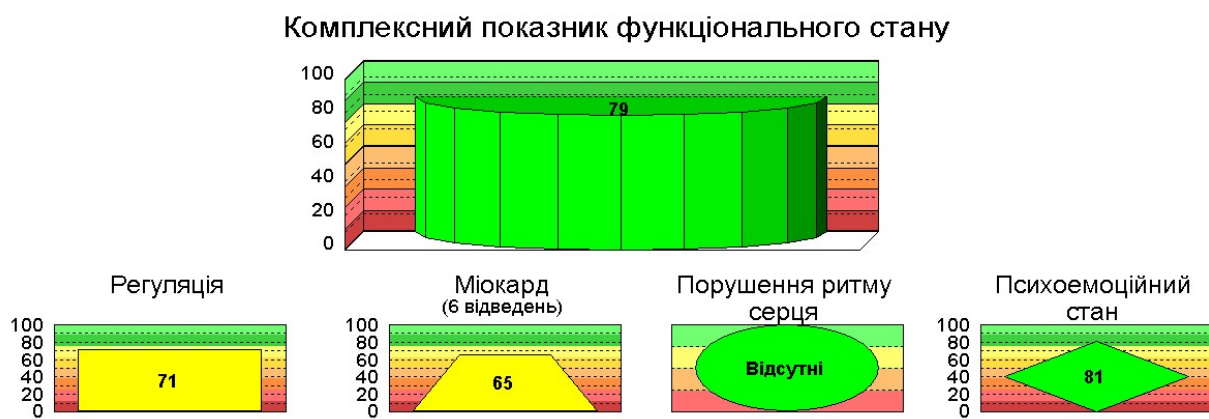


Рисунок 3.21 – Оцінка комплексного показника стану сигналів під номером 121 за допомогою програми Cardiol

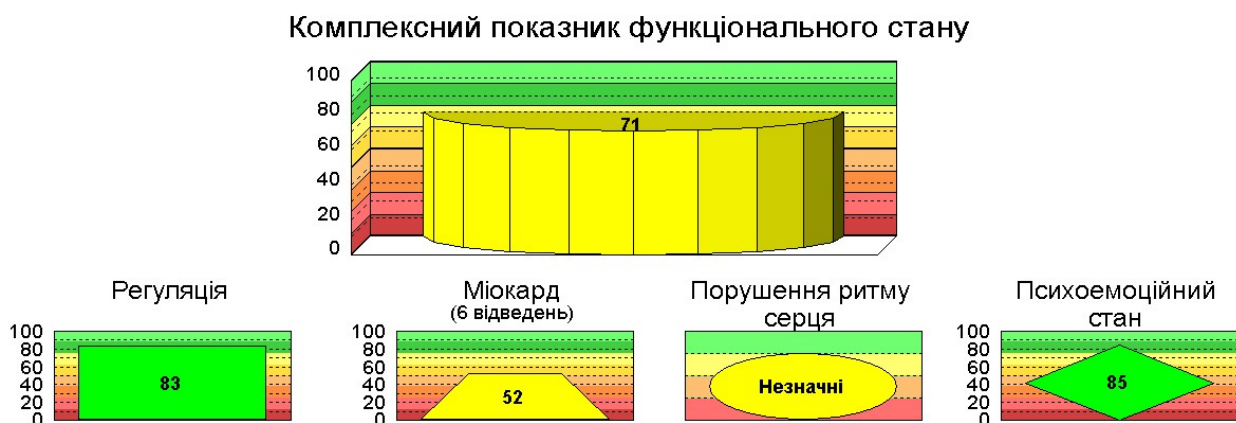


Рисунок 3.22 – Оцінка комплексного показника стану сигналів під номером 221 за допомогою програми Cardiol

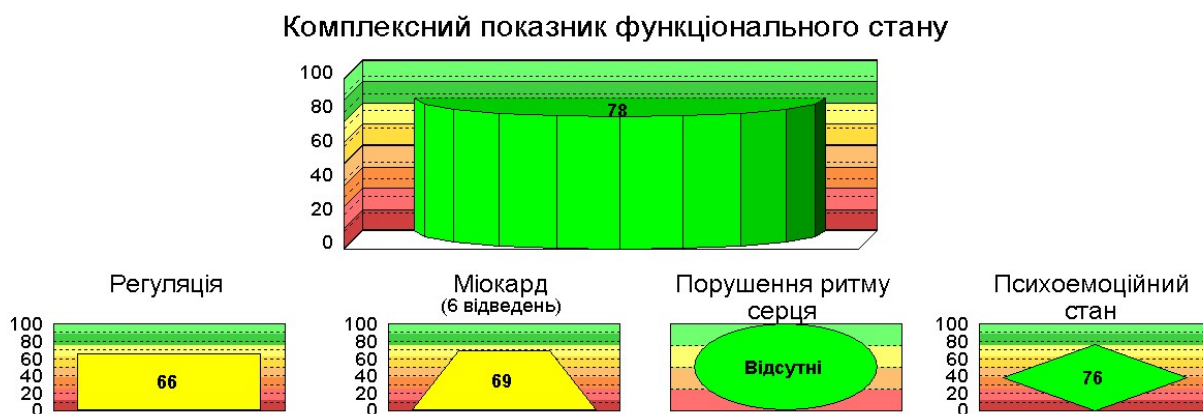


Рисунок 3.23 – Оцінка комплексного показника стану сигналів під номером 4 за допомогою програми Cardiol

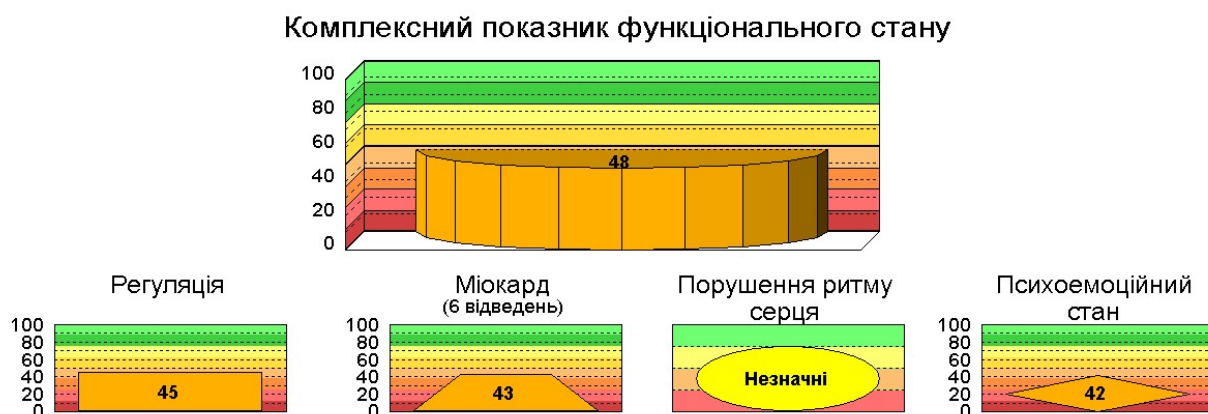


Рисунок 3.24 – Оцінка комплексного показника стану сигналів під номером 65 за допомогою програми Cardiol

На рисунках 3.12 – 3.23 представлені оцінки комплексного показника стану всіх пацієнтів, що були визначені алгоритмом як аномалії. З цих зображень видно, що всі викиди були пов'язані з погіршенням показників, які відповідають за стан міокарду. Це може бути як комплексне порушення (рисунки 3.14, 3.15, 3.17), так і порушення в одній шкалі показника (рисунки 3.13, 3.16), або двох. За результатами спілкування з лікарем було підтверджено, що ці пацієнти, дійсно мають деякі відхилення від норми. Найцікавіше, що сигнали з номерами 51, 52, 104 та 221 мають виключну особливість – як пізніше виявилось, вони не відносяться до

основної вікової категорії, представлений в даній базі, вони значно старші (одному з цих пацієнтів на момент вимірювань було близько 46 років). Це 4 з 6 подібних людей, наявних у вхідних даних. З цього висновок, що дослідження окремої вікової категорії має виконуватись окремо, оскільки поняття норми для неї сильно відрізняється. Сигнал під номером 121 був визначений як аномальний, тому що має помилковий показник значень амплітуди та тривалості зубця Р (тут вони дорівнюють нулю). Це може означати те, що алгоритм надає даним показникам велику вагу, тому вони сильно впливають на результат. Виходячи з цього було виконано дослід 5.1. В ньому алгоритму було задано параметр кількості використаних ознак. Тобто метод обирає найбільш вагомі параметри та виконує пошук аномалій за ними. Таким чином були внесені до норми сигнали з ідентифікаційними номерами 4, 108, 121, 221, які не мають виразних відхилень, як в тих, що залишились.

При поверхневому огляді було виділено одного пацієнта, показники якого програмою Cardiol оцінюються доволі погано (рисунок 3.24). Він не потрапив до списку аномалій в останніх дослідях, але потрапляв, коли параметрів на вході алгоритму було значно менше. Даний випадок говорить про наступне:

1. Кількості пацієнтів не достатньо для аналізу векторів високої розмірності. Тобто область норми розмита на стільки, що даний пацієнт вважається цілком нормальним на фоні інших;

2. Алгоритм враховує відношення показників один до одного. Тобто рівномірне відхилення усіх показників вважається нормою. Як видно з рисунку 3.24 показники знижені доволі рівномірно і слабо відрізняються один від одного, на відмінну від показників на рисунку 3.14, де присутнє явно виражене відхилення однієї частини параметрів, що відповідають за психоемоційний стан. Також це припущення можна використати відносно інших сигналів, оскільки в них є відхилення окремих комплексів показників відносно інших, тому вони і потрапили до списку аномалій.

3.4 Висновки до розділу 3

В результаті виконаної роботи було розглянуті наявні дані, створено функції для підготовки даних до експерименту (додаток А) та коду робочої програми, який розв'язує задачу дипломної роботи (додаток Б). З допомогою отриманого коду визначається деякий об'єм клінічно здорових людей, відносно якого можна визначити деяку «величину здоров'я серцево-судинної системи» за наявними показниками. В результатах наведені таблиці зі значенням відхилення пацієнта від норми. За деякими припущеннями дані значення можна вважати відстанню до ідеального стану, збірним чисельним показником здоров'я. Був отриманий список пацієнтів, дослідження яких, можливо, варто розширити, оскільки вони відрізняються від основної маси. Було визначено, що вік грає велику роль у визначенні норми. Припущення, наведені в роботах [10] та [12], а саме про необхідність аналізу SAECG для визначення стану міокарду та його функціональності після інфаркту були підтверджені (більша частина отриманих аномалій мають відхилення за показниками, пов'язаними саме з міокардом).

Створений підхід в аналізі ЕКГ потребує деяких правок:

1. Розширення та уточнення вхідних даних з метою створення більш чіткої області норми для використовуваного алгоритму;
2. Визначення найбільш вагомих параметрів норми та штучне збільшення їх ваги для алгоритму.

В теорії, наведені вище правки мають збільшити точність методу у визначенні аномалій для наявних даних. Варто зауважити, що задача, розглянута в даній роботі полягає у визначенні наявних аномалій за відсутності готового розв'язку, тобто виконуючи дану роботу було відомо лише те, що всі представлені у базі пацієнти клінічно здорові, а необхідно серед них визначити «менш здорових» та «більш здорових» не знаючи цей розподіл наперед.

ВИСНОВКИ

В даній дипломній роботі були розглянуті методи електрокардіографії. Одним з таких методів є аналіз SAEKG. Головною особливістю даного підходу є стійкість до шумів, це дозволяє аналізувати малоамплітудні сигнали, які на повному сигналі ЕКГ можуть сильно спотворюватись. Даний вид аналізу найчастіше використовують для відстеження стану міокарду після інфаркту, оскільки саме в діагностиці діяльності міокарду необхідно аналізувати малоамплітудні сигнали.

В даній роботі запропоновано використовувати SAEKG для визначення нормального стану серцево-судинної системи з допомогою методів машинного навчання. Було обрано метод Isolation Forest, який дозволяє знаходити аномальні значення серед інших, тобто вирізняти невідповідність показників окремого пацієнта групі людей зі схожими параметрами (вік, вага, загальний стан здоров'я, рід діяльності). Всі розглянуті сигнали були отримані з приймальної комісії військового училища. Даний факт створює наступні твердження:

1. Вікова категорія пацієнтів 17-19 років;
2. Всі оглянуті курсанти клінічно здорові, допущенні до навчання, фізичної і тактичної підготовки.

Оскільки всі сигнали відносять здорових людей, а необхідно визначити тих, хто чимось відрізняється, то необхідно зрозуміти «на скільки вони здорові», тобто визначити «величину здоров'я». Обраний метод машинного навчання (Isolation Forest) дозволяє порівнювати дані між собою, не задаючи початкових умов або вибірку значень, які точно є нормою, тобто метод сам визначає норму на основі вхідних даних. Для кожного вектору значень визначається відстань до визначеної методом норми, що можна вважати шуканою «величиною здоров'я».

В результаті поставленого експерименту був отриманий список пацієнтів, які відрізняються від основної маси. Виходячи з аналізу програми Cardiol, дані пацієнти мають відхилення з показниками, що відповідають за стан міокарду, що

знову підтверджує вірність методу аналізу SAECG після інфаркту. Після огляду результатів, лікар підтвердив, що дані пацієнти дійсно чимось відрізняються. До того ж, частина з них не є курсантами. До використаної бази даних потрапили офіцери училища, співробітники шпиталю, які не входять у вікову категорію, представлену в більшості. Це люди, значно старші. Цікаво те, що алгоритм визначив 4 з 6 таких випадків. Це означає, що для аналізу іншої вікової категорії необхідно використовувати інші норми. Інші пацієнти, що були визначені як аномалії, можливо, потраплять на більш прискіпливе дослідження, оскільки мають виразні відхилення деяких окремих параметрів сигналу ЕКГ.

В результаті роботи можна запропонувати деякі покращення даного підходу в майбутньому:

1. Уточнення бази даних, щоб уникнути ситуації, подібної тій, що виникла в роботі з іншою віковою категорією;
2. Збільшити кількість сигналів для більш точного визначення норми, бажано використовувати сигнали, виміряні не більш як 5 років тому, оскільки поняття норми досить швидко змінюється;
3. Визначити найвагоміші параметри для використовуваного методу;
4. Допрацювати алгоритм пошуку параметрів ще на етапі створення бази даних, оскільки деякі значення були визначені не адекватно і не відомо скільки ще ознак зазнали тих, чи інших спотворень.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Хэмптон Дж. Р. Основы ЭКГ: пер. с англ. — М.: Мед. лит., 2006 — 224 с., ил.
2. Воробьев Л.В. ЭКГ АНАЛИЗ СЕРДЕЧНОЙ ДЕЯТЕЛЬНОСТИ ЗДОРОВОГО ЧЕЛОВЕКА // Международный журнал прикладных и фундаментальных исследований. — 2016. — № 10-4. — С. 549-553
3. Ilya Chaikovsky. Electrocardiogram scoring beyond the routine analysis: subtle changes matters.
4. Cooper J. K. Electrocardiography 100 years ago. Origins, pioneers, and contributors (англ.) // N Engl J Med : journal. — 1986. — Vol. 315, no. 7. — P. 461—464. — doi:10.1056/NEJM198608143150721. — PMID 3526152.
5. Продукція та відомості про компанію ЮТАС.
Режим доступу: <http://www.utasco.com/>
6. Презентація прототипу приладу для можливих інвесторів в 2015 році.
Режим доступу: <https://startup.ua/news/cardiomo-na-strazhe-zdorovya-vashikh-blizkikh.html>
7. Онлайн лекція Expectation versus reality in biomedical signal processing. Спікер : Артем Агарков. Algorithms Engineer в Cardiomo.
Режим доступу: <https://www.youtube.com/watch?v=pxqjebp5hq4>
8. Kim, Dae-Hyeong & Lu, Nanshu & Ma, Rui & Kim, Yun-Soung & Kim, Rak-Hwan & Wang, Shuodao & Wu, Jian & Won, Sang & Tao, Hu & Islam, Ahmad & Yu, Ki & Kim, Tae-il & Chowdhury, Raed & Ying, Ming & Xu, Lizhi & li, Ming & Chung, Hyun-joong & Keum, Hohyun & McCormick, Martin & Rogers, John. (2011). Epidermal Electronics. Science (New York, N.Y.). 333. 838-43. 10.1126/science.1206157.
9. J. BERBARI, M.S., RALPH LAZZARA, M.D., PHI SAMET, M.D., AN BENJAMIN J. SCHERLAG, PH.D. Noninvasive Technique for Detection of Electrical Activity During the P-R Segment

- Режим доступу: <https://www.ahajournals.org/doi/pdf/10.1161/01.CIR.48.5.1005>
10. Signal-averaged electrocardiography: Past, present, and future. Konstantinos A. Gatzoulis, Petros Arsenos, Konstantinos Trachanas, Polychronis Dilaveris, Christos Antoniou, Dimitris Tsiachris, Skevos Sideris, Theofilos M. Kolettis, Dimitrios Tousoulis *J Arrhythm*. 2018 Jun; 34(3): 222–229. Published online 2018 May 28. doi: 10.1002/joa3.12062 PMCID: PMC6010001
 11. Mewton N, Strauss DG, Rizzi P, et al. Screening for cardiac magnetic resonance scar features by 12-Lead ECG, in patients with preserved ejection fraction. *Ann Noninvasive Electrocardiol*. 2016;21:49–59
 12. G. Breithardt, T. Wichter, T. Fetsch, T. Seifert, M. Borggrefe, M. Shenasa, L. Reinhardt, The signal-averaged ECG: Time-domain analysis, *European Heart Journal*, Volume 14, Issue suppl_E, 1 September 1993, Pages 27–32.
 13. Jongman JK, Zaidi A, Muggenthaler M, Sharma S. Relationship between echocardiographic right-ventricular dimensions and signal-averaged electrocardiogram abnormalities in endurance athletes. *Europace*. 2015;17:1441–8
 14. Zaidi A, Sheikh N, Jongman JK, et al. Clinical differentiation between physiological remodeling and arrhythmogenic right ventricular cardiomyopathy in athletes with marked echocardiographic repolarization anomalies. *J Am Coll Cardiol*. 2015;65:2702–11.
 15. Стаття на науково-популярному порталі. Искусственный интеллект: вчера, сегодня, завтра
Режим доступу: <https://intellect.icu/iskusstvennyj-intellekt-vchera-segodnya-zavtra-198>
 16. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*
 17. Весенков Данила Валентинович. Методы обучения искусственных нейронных сетей.
 18. Buchanan, B. G., & Feigenbaum, E. A. (1978). Dendral and meta-dendral: Their applications dimension.

19. Swartout, W. R. (1985). Rule-based expert systems: The mycin experiments of the stanford heuristic programming project.
20. Aikins, J. S., Kunz, J. C., Shortliffe, E. H., & Fallat, R. J. (1983). PUFF: An expert system for interpretation of pulmonary function data
21. Mincholé, A., Camps, J., Lyon, A., & Rodríguez, B. (2019). Machine learning in the electrocardiogram. *Journal of Electrocardiology*.
22. Özbay, Y., Ceylan, R., & Karlik, B. (2006). A fuzzy clustering neural network architecture for classification of ECG arrhythmias. *Computers in Biology and Medicine*, 36(4), 376–388.
23. Опис та приклад функції ієрархічної кластеризації в бібліотеці scikit-learn мови програмування Python
 Режим доступу: https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html
24. Mahalanobis, Prasanta Chandra. On the generalised distance in statistics // *Proceedings of the National Institute of Sciences of India*. — 1936. — Т. 2, № 1. — С. 49—55.
25. Nello Cristianini, John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. — Cambridge University Press, 2000.
26. Liu, Fei Tony; Ting, Kai Ming; Zhou, Zhi-Hua (December 2008). "Isolation Forest". *2008 Eighth IEEE International Conference on Data Mining*: 413–422.
27. Опис функцій та параметрів методу Isolation Forest в бібліотеці на мові програмування Python.
 Режим доступу: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html>

Додаток А

```

# data_preparation.py
import xml.etree.ElementTree as ET
import os
import numpy as np

def opening(dir):

    # directory that consist data files only

    # The data structure is as follows:
    # ID1 : [lead : data, param : param_value for 6 leads]
    # ...
    # ...
    # IDN : [lead : data, param : param_value for 6 leads]

    os.chdir(dir)
    global_dictionary = {}
    file_names = os.listdir(path=".") # names of data files

    for name in file_names:
        tree = ET.parse(name)
        root = tree.getroot()

        channels = root.find('INTERVAL').findall('CHANNEL')
        param_tag = []

        a = channels[0].find('PARAMETERS')
        if a is not None:
            for el in a:
                param_tag.append(el.tag)

        data_dictionary = {}
        value = []
        for ch in channels:

            param_text = []

            a = ch.find('PARAMETERS')
            if a is not None:
                for el in a:

```

```

        param_text.append(el.text)
        value.append(param_text)

    data_dictionary[ch.find('ID').text] = list(map(int, ch.find('DATA').text.split(' ')))

    global_dictionary[name] = data_dictionary

for i, el in enumerate(param_tag):
    if i == 0:
        data_dictionary[el] = [value[j][i] for j in range(6)]
    else:
        data_dictionary[el] = [int(value[j][i]) for j in range(6)]

param_names = list(data_dictionary.keys())    # names of lead and parameters
return (global_dictionary, file_names, param_names)

def array_formation (initial_DATA, lead_nums, feature_names):

#lead_nums - array that consist numbers of leads from 0 to 5
#feature_names - names of parameters that needs

# The data structure is as follows:
# [[feature 1, feature 2, ..., feature N for lead 1, ..., feature 1, feature 2, ..., feature N for
lead N] for ID1
# [.....] for ID2
# [.....] ...
# [.....]] for IDN

    data_array = []
    for id in list(initial_DATA.keys()):
        X = []
        for num in lead_nums:
            for i in range(len(feature_names)):
                X.append(initial_DATA[id][feature_names[i]][num])
        data_array.append(X)
    return(np.array(data_array))

```

Додаток Б

```

import matplotlib.pyplot as plt
import data_preparation
from sklearn.ensemble import IsolationForest
import numpy as np

[DATA, ID, param_names] = data_preparation.opening('C:\\Education\\diploma\\data')
param_names.pop(10)

#[Lead I, Lead II, Lead III, Lead aVR, Lead aVL,
# Lead aVF, PC1, PC2, PC3, [PC],
# -----, PR_DURATION, P_DURATION, P_AMPLITUDE, P_NOTCHED,
# P1_DURATION, P1_AMPLITUDE, P_AREA, QRS_DURATION,
Q_DURATION,
# Q_AMPLITUDE, R_DURATION, R_AMPLITUDE, S_DURATION,
S_AMPLITUDE,
# R1_DURATION, R1_AMPLITUDE, S1_DURATION, S1_AMPLITUDE,
T_AMPLITUDE,
# T1_AMPLITUDE, J_AMPLITUDE, ST_MID_AMPLITUDE,
ST_END_AMPLITUDE, QRS_AREA,
# TBEG_T_AREA, T_TEND_AREA, TBEG_T_MAX_DER, T_TEND_MAX_DER,
QR_DURATION,
# TR_DURATION, T1R_DURATION, JT_DURATION, JTA_DURATION,
TA_TEND_DURATION]
n = [0,1,2,3,4,5]
features = [PR_DURATION, P_DURATION, P_AMPLITUDE, P_NOTCHED,
P1_DURATION, P1_AMPLITUDE, P_AREA, QRS_DURATION,
Q_DURATION,
Q_AMPLITUDE, R_DURATION, R_AMPLITUDE, S_DURATION,
S_AMPLITUDE,
R1_DURATION, R1_AMPLITUDE, S1_DURATION, S1_AMPLITUDE,
T_AMPLITUDE,
T1_AMPLITUDE, J_AMPLITUDE, ST_MID_AMPLITUDE,
ST_END_AMPLITUDE, QRS_AREA,
TBEG_T_AREA, T_TEND_AREA, TBEG_T_MAX_DER, T_TEND_MAX_DER,
QR_DURATION,
TR_DURATION, T1R_DURATION, JT_DURATION, JTA_DURATION,
TA_TEND_DURATION]
using_array = data_preparation.array_formation(DATA, n, features)

plt.figure()

```

```

xlab = "Тривалість, мСек"
ylab = 'Амплітуда, мкВ'
plt.subplot(1, 2, 1); plt.title (features[0]); plt.xlabel(xlab); plt.ylabel("Кількість, шт")
plt.hist(using_array[:, 0], bins=30, histtype='bar')
plt.subplot(1, 2, 2); plt.title (features[6]); plt.xlabel(ylab); plt.ylabel("Кількість, шт")
plt.hist(using_array[:, 6], bins=30, histtype='bar')

# iForest training
#pattern1
#pattern2 n_estimators=50, contamination=0.1
#pattern3 n_estimators=200, contamination=0.05

clf = IsolationForest(max_features=0.1, n_estimators=200)
y_pred = clf.fit_predict(using_array)
dec = clf.decision_function(using_array)

# Outliers identification
outliers = []
for i, el in enumerate(y_pred):
    if el == -1:
        outliers.append(i)
sort_dict = {}
for el in outliers:
    sort_dict[dec[el]] = ID[el]
for key in sorted(sort_dict):
    print(sort_dict[key])
for key in sorted(sort_dict):
    print("{:.6f}".format(key))

plt.figure(); plt.xlabel(xlab); plt.ylabel(ylab)
otv = 1
print(len(features))
b1 = plt.scatter(using_array[:, 1+len(features)*otv], using_array[:, 2+len(features)*otv],
c='white', s=10, edgecolor='k')
b2 = plt.scatter(using_array[outliers, 1+len(features)*otv], using_array[outliers,
2+len(features)*otv], c='red', s=40, edgecolor='k')
plt.show()

```